About Phase: Synthetic Aperture Radar and the Phase Retrieval Problem

THESIS

Aaron A. Nelson, 2d Lt, USAF

AFIT-ENC-14-M-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, the Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENC-14-M-03

About Phase: Synthetic Aperture Radar and the Phase Retrieval Problem

THESIS

Presented to the Faculty

Department of Mathematics and Statistics

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science

Aaron A. Nelson, 2d Lt, USAF, BS

March 2014

AFIT-ENC-14-M-03

About Phase: Synthetic Aperture Radar and the Phase Retrieval Problem

Aaron A. Nelson, 2d Lt, USAF, BS

Approved:

//signed//	14 March 2014
Dustin G. Mixon, Capt, USAF, PhD (Chair)	Date
//signed//	14 March 2014
Benjamin F. Akers, PhD (Member)	Date
//signed//	14 March 2014
Jesse D. Peterson, Capt, USAF, PhD (Member)	Date

AFIT-ENC-14-M-03

Abstract

Synthetic aperture radar (SAR) uses relative motion to produce fine resolution images from microwave frequencies and is a useful tool for regular monitoring and mapping applications. Unfortunately, if target distance is estimated poorly, then phase errors are incurred in the data, producing a blurry reconstruction of the image. In this thesis, we introduce a new multistatic methodology for determining these phase errors from interferometry-inspired combinations of signals. To motivate this, we first consider a more general problem called phase retrieval, in which a signal is reconstructed from linear measurements whose phases are either unreliable or unavailable. We make significant theoretical progress on the phase retrieval problem, to include characterizing injectivity in the complex case, devising the theory of almost injectivity, and performing a stability analysis. We then apply certain ideas from phase retrieval to resolve phase errors in SAR. Specifically, we use bistatic techniques to measure relative phases, and then we apply a graph-theoretic phase retrieval algorithm to recover the phase errors. We conclude by devising an image reconstruction procedure based on this algorithm, and we provide simulations that demonstrate stability to noise.

Keywords: Synthetic aperture radar, phase retrieval, angular synchronization, phase errors, circulant graphs, informationally complete, quantum mechanics, unit norm tight frames, computational complexity, Cramer-Rao lower bound

Acknowledgements

I would like to personally thank Matthew Fickus and Jesse Peterson, both of whom significantly influenced the development of this thesis with their thoughts and insights. Others deserving of recognition include Afonso Bandeira, Jameson Cahill and Yang Wang for their contributions to the ideas presented in Chapters II, III and IV. I extend my sincerest thanks to my research advisor, Dustin Mixon, whose constant motivation and drive were pivotal in the completion of this document; his attitude is infectious and his patience truly endless. Moreover, his constant guidance and instruction have made my time at AFIT one of the most rewarding experiences of my academic career. I am grateful to have had the opportunity to work with him and, most importantly, learn from him.

This research was supported by NSF Grant No. DMS-1321779.

SAR image of Pentagon courtesy of Sandia National Laboratories [83]

Aaron A. Nelson, 2d Lt, USAF

Table of Contents

List of Figures

About Phase: Synthetic Aperture Radar and the Phase Retrieval Problem

I. Introduction

1.1 Synthetic aperture radar

Synthetic aperture radar (SAR) is a form of radar that uses relative motion to produce fine resolution images from microwave signals. The usefulness of SAR stems from its ability to overcome the shortcomings of competing remote imaging systems. For instance, its day-or-night and all-weather capabilities give SAR an advantage over both optical cameras and infrared imagers while maintaining comparable spatial resolution [48]. As such, SAR is a particularly useful tool for regular monitoring and mapping applications, some of which include the following:

Reconnaissance and surveillance. SAR imaging enables constant reconnaissance and surveillance, as it can operate at any time of day and in all weather conditions, while offering sufficient resolution to distinguish terrain features and identify man-made targets. Even moving targets may be identified, and so SAR is capable of monitoring traffic patterns or tracking the movement of personnel and vehicles [2,74].

Topography. With the help of certain interferometric techniques, SAR can be used to create accurate topographic maps and surface profiles. The extremely high resolution of these techniques also enables detection of sudden seismic activity and even volcanic bulging prior to the eruption of volcanoes [2, 48, 71].

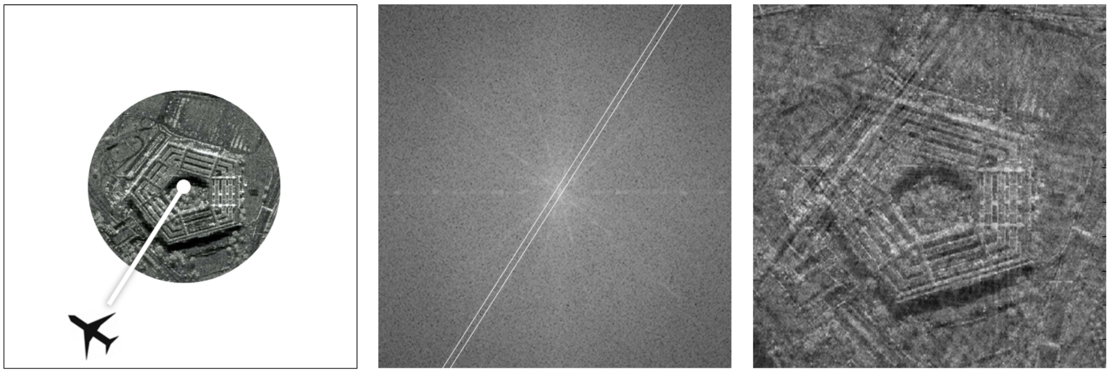
Navigation and guidance. All-weather, autonomous navigation and guidance may be accomplished using SAR by periodically imaging the surrounding terrain and comparing to a stored reference image. This comparison then provides a means for navigation update, and can even be used to accurately guide aircraft or munitions to a target [74].

Foliage and ground penetration. Due to its use of microwave frequencies, SAR offers the capability of penetrating optically opaque materials, such as foliage and topsoil. Thus, SAR enables monitoring of activity normally hidden by trees, brush, or similar ground cover. Depending on soil conditions, SAR is also capable of imaging underground targets of sufficient size at depths of up to several meters [2, 74].

Environmental monitoring. SAR is particularly sensitive to the dielectric properties of materials, making it useful for monitoring the condition of vegetation. Thus, it is an important agricultural and environmental tool, capable of accurately monitoring crop characteristics, soil moisture levels, deforestation, ice flows, and oil spills. In particular, SAR is effective at detecting oil spills over open water due to certain backscatter effects [2, 74].

SAR works by implementing a moving radar platform that repeatedly transmits a certain type of microwave signal and records the return signal reflected by the scene of interest. Typical platforms used for SAR imaging include aircraft and satellite, although each platform presents its own challenges during reconstruction [48]. Since the radar source is in motion (relative to the target), repeatedly imaging a scene provides information from a continuously changing perspective, and it is precisely this introduction of perspective to the system that enables image reconstruction with increased resolution.

In airborne spotlight-mode SAR, appropriate assumptions regarding the transmitted signal (e.g., assumptions relating its frequencies to the speed of light) along with assumptions about the scene (e.g., that elevations are relatively constant, so the desired image is simply a function over $\mathbb{R}^2$) enable the return signal to be interpreted in terms of the Fourier transform of the target image. Indeed, under these assumptions, the signal reflected back to the radar source is the transmitted signal multiplied by a unit-modulus phase factor ω, and pointwise multiplied by a predictably modulated version of a one-dimensional slice of the Fourier transform of the desired reflectivity function $\rho \colon \mathbb{R}^2 \to \mathbb{R}$, which describes certain electromagnetic



(a) (b) (c)

Figure 1: (a) Classical airborne spotlight-mode synthetic aperture radar (SAR). The aircraft transmits a signal and receives a version of that signal which encodes a portion of the Fourier transform of the desired image. (b) Based on the aircraft's current position, it obtains the depicted slice of the Fourier transform. (c) After obtaining a range of slices of the image's Fourier transform, the slices are interpolated before inverting the Fourier transform. Unfortunately, if the distance between the aircraft and the scene of interest is estimated poorly, then a phase error is incurred in the corresponding slice. Different phase errors for different slices accumulate to produce a blurry reconstruction of the desired image. Such phase errors are typically estimated and removed using various post-processing techniques, and while these tend to work rather well, they are often ad hoc, requiring additional assumptions about the target scene, and they sometimes fail unexpectedly. One of the contributions of this thesis is the introduction of a new multistatic methodology for determining these phase errors from interferometry-inspired combinations of signals.

characteristics of the scene [48] (cf. Fact 5.1 in this thesis). Specifically, if one transmits and receives a signal from a point (x, y) to image a scene which is centered at the origin, then the received signal encodes the portion of the scene's two-dimensional Fourier transform $F\rho$ that lies on the line which passes through both (x, y) and the origin (see Figure 1(a) and (b) for an illustration). In practice, such portions are interpolated to reconstruct the entire two-dimensional Fourier transform, from which the image may be easily obtained [48]. Unfortunately, an issue arises when using this approach, namely, uncertainty in the target distance (that is, the distance from the radar source to the target scene). Indeed, the phase factor ω which appears in the received signal is a sensitive function of target distance. Furthermore, small fluctuations in this distance are quite common due to factors such as aircraft performance, weather, wind, and pilot skill [18]. Overall, any noise in the estimates of these distances creates *phase errors* in the recorded signals, and since the phase error will be different for each slice of the Fourier transform, the image becomes distorted when taking the inverse Fourier transform.

The effect of phase errors is pointwise multiplication in the Fourier domain, meaning the desired image is blurred in the spatial domain, typically enough so that objects of interest within the target scene are indiscernible (see Figure 1(c), for example). Although there are methods for dealing with phase errors during post-processing (e.g., autofocus algorithms [63]), and many of these certainly produce outstanding results, it is desirable to eliminate the problem prior to image reconstruction. Indeed, many algorithms for correcting phase errors use ad hoc techniques, require further assumptions on the target scene, and may fail unexpectedly [37, 44, 54, 79].

It is reasonable to expect that the phase error problem in SAR could have a more systematic solution if only additional signal data were available for analysis. The desire for more information motivates the use of *multistatic* radar systems, in which multiple radar sources, separated by distances comparable to any single target-to-source distance, are capable of both transmitting and receiving signals reflected by a common target. These systems enable multiple measurements to be taken from varying perspectives, and when combined, these additional measurements often improve resolution and even combat some of the weaknesses of monostatic radar systems [71]. To date, multistatic techniques have been used for various applications, such as tracking and triangulation, using both stationary and mobile radar sources [59, 64, 71]. For instance, antenna "swarms" are a common airborne application of multistatic radar for target tracking in which multiple radar platforms with independent flight paths are capable of both transmitting and receiving signals to and from a (possibly moving) target [13, 36, 71]. One way of realizing these swarms is to mount radar receivers on a team of remotely piloted aircraft (RPAs); in such multistatic systems, a common radar source transmits a signal to a target while each RPA records time-delay and Doppler measurements of the reflected signal which, when combined, provide target tracking that has been shown to outperform traditional, static radar arrays [13,33]. Even passive sources like radio and television broadcast signals can be incorporated in a multistatic system [15, 59, 64].

The techniques of monostatic SAR described earlier can be naturally extended to the bistatic and multistatic settings [48]. As we will demonstrate, the introduction of additional radar transmitters and receivers allows one to observe interferometry-inspired combinations of the phase errors we seek to remove. We will then apply ideas from a related, well-studied problem called *phase retrieval* to estimate the phase errors from these combinations.

1.2 The phase retrieval problem

The phase error problem in SAR can be viewed as an instance of a more general problem called phase retrieval, in which one attempts to reconstruct a signal when phase information is either unreliable (as in the case of SAR) or completely lost during some linear measurement process. Indeed, given slices of the Fourier transform, each multiplied by a different unknown phase factor, one can simply ignore any phase information by taking pointwise absolute values, effectively reducing the phase error problem in SAR to the most common problem in phase retrieval: recover an image from the pointwise absolute value of its Fourier transform. This reduction implies that any method of phase retrieval is also a solution to the phase error problem.

We note that phase retrieval is interesting in its own right, as it has many applications other than SAR:

Coherent diffractive imaging. A common technique for imaging a nanoscale object is to strike it with a highly coherent beam of X-rays and record the resultant diffraction pattern using a photon counting device. This diffraction pattern is the Fourier transform of the material density profile of the object. However, counting photons only provides the intensity of the diffraction pattern, and so recovering the image first requires obtaining the lost phase information via phase retrieval [19, 60, 76, 78].

Optics. This application enjoys various instances of phase retrieval:

(i) In astronomy, imaging celestial objects like stars using a lens-based optical system requires computing the associated pupil distribution (i.e., the distribution of light that is allowed to exit the optical system). Since such systems only detect the pointwise absolute value of the Fourier transform of the pupil function, one must first recover the phases before building an image of the object [91].

(ii) When producing a high-resolution image of a radiating object, certain interferometric techniques can be used to approximate the object's spatial coherence function, which is the Fourier transform of the object map (i.e., the spatial intensity of the radiation). Unfortunately, the phase of this function is quite difficult (and often impossible) to estimate accurately, and so is typically discarded in favor of estimation by phase retrieval [40].

(iii) Soon after NASA launched its Hubble Space Telescope, it was discovered that its primary mirror suffered from a large spherical aberration (i.e., phase errors resulting from light striking the mirror near its edge). To determine the proper correction, the extent of the aberration was established by constructing the pupil function from the associated point spread function (which measures the intensity of the Fourier transform of the pupil function) using phase retrieval [55].

Quantum state tomography. When measuring a pure quantum state using a positive operator–valued measure (POVM) of rank-1 elements, the distribution of the random outcome of the measurement can be expressed in terms of the state's intensity measurements with the Parseval frame elements which generate the POVM. Since repeated measurements of the state produce an empirical estimate of this distribution, phase retrieval can be used to identify the state [56, 57, 62].

Speech processing. In signal processing for speech applications, a common method of denoising is to take the short-time Fourier transform (STFT) and perform a smoothing operation on the magnitudes of the coefficients. Instead of inverting the STFT using the (noisy) unaltered phases of the coefficients, one can recover the denoised version of the signal by first discarding the phases and then reconstructing with phase retrieval [8, 85].

Although there are many applications of phase retrieval, the task is often impossible. For instance, intensity measurements with the identity basis effectively discard the phase information of a signal's entries, and so this measurement process is not at all injective; similarly, the power spectrum discards the phases of Fourier coefficients. This fact has led many researchers to invoke a priori knowledge of the desired signal, since intensity measurements might be injective when restricted to a smaller signal class. This is frequently the case in optics applications, since the pupil distribution is only supported within the aperture of the optical system; the resultant compact-support constraint is often sufficient to make the intensity measurement mapping injective. The introduction of such information has led to various ad hoc phase retrieval algorithms, and while some have found success (e.g., in correcting the Hubble Space Telescope), such algorithms often fail to work unexpectedly. (The situation is not unlike the state of the art for correcting phase errors in SAR.) Overall, algorithms produced in this way typically lack practical performance guarantees.

Thankfully, there is an alternative approach to phase retrieval, as introduced in 2006 by Balan, Casazza and Edidin [8]: Instead of restricting to a smaller signal class, seek injectivity by designing a larger ensemble of measurement vectors. (This approach is an underlying theme throughout this thesis.) Unbeknownst to Balan et al. at the time, the quantum mechanics community was already familiar with this idea (for quantum state tomography [56, 57]), but presenting the idea to the signal processing community led to a flurry of research in search of practical phase retrieval guarantees [3, 5, 7, 9, 24–27, 43, 49, 88, 90].

At this point, it is helpful to introduce some notation. Given a collection of measurement vectors $\Phi = \{\varphi_n\}_{n=1}^N$ in $V = \mathbb{R}^M$ or $\mathbb{C}^M$, which we identify with the $M \times N$ matrix whose columns form the collection, we consider the *intensity measurement process* defined by

$$(\mathcal{A}(x))(n) := |\langle x, \varphi_n \rangle|^2.$$

For example, in the case of phase retrieval with the Fourier transform, each φ_n is a complex sinusoid and Φ^* is the Fourier transform. Note that $\mathcal{A}(x) = \mathcal{A}(y)$ whenever $y = cx$ for some scalar c of unit-modulus. As such, the mapping $\mathcal{A}\colon V \to \mathbb{R}^N$ is not injective. To resolve this (technical) issue, throughout this thesis we consider sets of the form V/S, where V is a vector space and S is a multiplicative subgroup of the field of scalars. By this notation, we mean to identify vectors $x, y \in V$ for which there exists a scalar $c \in S$ such that $y = cx$; we write $y \equiv x \bmod S$ to convey this identification. Most (but not all) of the time, V/S is either $\mathbb{R}^M/\{\pm 1\}$ or $\mathbb{C}^M/\mathbb{T}$ (here, $\mathbb{T}$ is the complex unit circle), and we view the intensity measurement process as a mapping $\mathcal{A}\colon V/S \to \mathbb{R}^N$; it is in this way that we will consider the measurement process to be injective or stable.

In order to perform phase retrieval successfully, we therefore seek to understand the properties of the measurement ensemble Φ that enable recovery of a signal x from measurements of the form $\mathcal{A}(x)$. This naturally leads to the following question:

The Phase Retrieval Problem. *What are necessary and sufficient conditions for efficient and stable recovery of a signal from its intensity measurements?*

As a noteworthy stride toward solving the phase retrieval problem, Candès, Strohmer and Voroninski [27] viewed intensity measurements as Hilbert-Schmidt inner products between rank-1 operators, and they applied certain intuition from compressed sensing to stably reconstruct the desired M-dimensional signal with semidefinite programming; similar alternatives and refinements have since emerged [24, 34, 43, 89]. Another alternative exploits the polarization identity to discern relative phases between certain intensity measurements; this method uses an expander graph along with a new algorithm called angular synchronization to quickly solve certain instances of phase retrieval, and it comes with a similar stability guarantee [3, 9]. One can also formulate phase retrieval in terms of MaxCut, and solvers for this formulation are equivalent to a popular solver (PhaseLift) for the matrix recovery formulation [88, 90]. In this same line of research, a new methodology for coherent diffractive imaging emerged [24]: Rather than attempting phase retrieval with possibly incomplete information taken from a single exposure, take multiple exposures of the same object using different diffraction gratings. Such a process is capable of producing complete information and is associated with provably efficient (and apparently stable) phase retrieval algorithms [9, 26]. This approach inspires the use of multistatic SAR in this thesis as a means of producing complete information for the phase error problem.

1.3 Overview

This thesis offers two main contributions: (i) we make significant theoretical progress on the phase retrieval problem, and (ii) we apply certain ideas from phase retrieval to resolve phase errors in synthetic aperture radar. We begin in Chapter II by examining what it means for an ensemble of intensity measurements to be injective. In particular, we discuss the characterization of injectivity in the real case as introduced by Balan, Casazza and Edidin [8], i.e., the *complement property*, and provide the first known characterization of injectivity in the complex case (Theorem 2.3). Next, we make a rather surprising identification: that intensity measurements are injective in the complex case precisely when the corresponding phase-only measurements are injective in some sense (Theorem 2.4). We then use this identification to prove the necessity of the complement property for injectivity (Theorem 2.6). Later, we formulate a conjecture that $4M-4$ intensity measurements are necessary and sufficient for injectivity in the complex case, as well as discuss the cases for which the conjecture is known to hold; we also prove several such cases. Specifically, for the proof of the case $M=3$ we introduce a new test for injectivity, which we then use to verify the injectivity of a certain quantum mechanics–inspired measurement ensemble, thereby suggesting a new refinement of Wright's conjecture from [87] (see Conjecture 2.14). The chapter concludes with an explicit construction of $4M-4$ intensity measurements which yield injectivity, the second known injective ensemble of this size (the first is due to Bodmann and Hammen [17]). Bodmann and Hammen [17] leverage the Dirichlet kernel and the Cayley map to prove injectivity of their ensemble, but it is unclear whether phase retrieval is algorithmically feasible from their ensemble. By contrast, for the ensemble in this thesis, we use basic ideas from harmonic analysis over cyclic groups to devise a corresponding phase retrieval algorithm, and we demonstrate injectivity in Theorem 2.20 by proving that the algorithm recovers any noiseless signal up to global phase.

In Chapter III, we devise a theory of ensembles for which the corresponding intensity measurements are "almost" injective, that is, are injective on a set of signals that is dense in $\mathbb{C}^M$. Here, we focus on the real case, meaning phase retrieval is up to a global sign factor $\omega = \pm 1$, and our approach is inspired by the characterization of injectivity in the real case by Balan, Casazza and Edidin [8]. After characterizing almost injectivity in the real case, we find a particularly satisfying sufficient condition for almost injectivity: that the ensemble of measurement vectors forms a unit norm tight frame with relatively prime dimensions (Theorem 3.7). Characterizing almost injectivity in the complex case remains an open problem. The chapter concludes with a discussion of the computational limits of phase retrieval, in which we consider algorithmic phase retrieval in the real case using $M+1$ almost injective intensity measurements. Specifically, we show that phase retrieval in this case is NP-hard by reduction from the subset sum problem (Theorem 3.9). The hardness of phase retrieval in this minimal case suggests a new problem for phase retrieval: What is the smallest C for which there exists a family of ensembles of size $CM + o(M)$ such that phase retrieval can be performed in polynomial time?

We devote Chapter IV to stability in phase retrieval. Here, we start by focusing on the real case, for which we give upper and lower Lipschitz bounds of the intensity measurement mapping in terms of singular values of submatrices of the measurement ensemble (Lemma 4.3 and Theorem 4.5); this suggests a new matrix condition called the *strong complement property*, which strengthens the complement property of Balan et al. [8] and bears some resemblance to the restricted isometry property of compressed sensing [23]. As we will discuss, our result corroborates the intuition that localized frames fail to yield stability. We then show that Gaussian random measurements satisfy the strong complement property with high probability (Theorem 4.7), which nicely complements certain results of Eldar and Mendelson [49]. In particular, we find an explicit, intuitive relation between the Lipschitz bounds and the number of intensity measurements per dimension (see Figure 3). Finally, we present results in both the real and complex cases using a stochastic noise model, much like Balan did for the real case in [5]; here, we leverage Cramer-Rao lower bounds to identify stability with stronger versions of the injectivity characterizations (see Theorems 4.8 and 4.10).

Chapter V finally returns to the phase error problem in synthetic aperture radar. By incorporating techniques from bistatic radar, we formulate the phase error problem in terms of relative phases, bearing some resemblance to those obtained from interferometric intensity measurements used for phase retrieval by Alexeev, Bandeira, Fickus and Mixon [3]. In particular, Alexeev et al. leverage an algorithm known as *angular synchronization* [84] to recover a set of phases from their relative phase measurements, which motivates a graph theoretic approach to the phase error problem in SAR. Using this approach, we then formulate phase error recovery as a feasibility problem, solutions to which are only unique up to a modulation and global phase. We conclude by constructing an algorithm that extracts phase errors from multistatic SAR data using certain graphs that can be obtained from particular arrangements of different numbers of aircraft. Our two-step image reconstruction algorithm first uses an iterative form of angular synchronization to determine the phase errors up to a modulation and a single global phase factor, and then maximizes the image's total variation to determine the appropriate modulation and phase factor. Simulations with random phase error data are provided, with which it is shown that the algorithm exhibits stability in terms of the number of cycles contained in the parent graph. In particular, the number of cycles is directly related to the number of aircraft used in the multistatic system, and the simulations suggest that the phase error problem can be solved using only a few aircraft (e.g., as few as five for a graph of 101 vertices).

We conclude in Chapter VI with some discussion and ideas for future work. For the record, the material presented in Chapters II, III, and IV also appears in three peer-reviewed publications. Sections 2.1 and 2.2, as well as Chapter IV and Appendix A have appeared in the proceedings of the 10th International Conference on Sampling Theory and Applications [10], and a journal version of the conference paper has been accepted for publication in Applied and Computational Harmonic Analysis [11]. Also, Section 2.3 and Chapter III appear in a journal article which has been accepted for publication in Linear Algebra and its Applications [52].

II. Injective intensity measurements and the $4M-4$ conjecture

An underlying theme in the phase retrieval problem is determining necessary and sufficient conditions for the intensity measurement process $\mathcal{A}$ to be injective. Indeed, injectivity ensures complete information for signal reconstruction, and so these conditions are important specifications for the intensity measurement process. Recall that, given a collection of measurement vectors $\Phi = \{\varphi_n\}_{n=1}^N$ in $V = \mathbb{R}^M$ or $\mathbb{C}^M$, the intensity measurement process $\mathcal{A}$ cannot be injective if viewed as a mapping from V into $\mathbb{R}^N$. For this reason, we identify vectors $x, y \in V$ for which there exists a scalar $c \in S$ such that $y = cx$, and we view the intensity measurement process as a mapping $\mathcal{A}\colon V/S \to \mathbb{R}^N$, where $S = \{\pm 1\}$ or $\mathbb{T}$. In this chapter, we will examine what it means for an ensemble of intensity measurements to be injective. We characterize injectivity in both the real and complex cases before focusing on injectivity with the absolute minimum number of intensity measurements. This leads to the conjecture that $4M-4$ intensity measurements are necessary and sufficient for injectivity in the complex case (Conjecture 2.9). The remainder of the chapter is dedicated to making progress on this conjecture, including a deterministic construction of $4M-4$ intensity measurements that yield injectivity.

2.1 Injectivity and the complement property

Phase retrieval is impossible without injective intensity measurements. As such, we desire necessary and sufficient conditions on the size of an ensemble of M-dimensional measurement vectors $\Phi = \{\varphi_n\}_{n=1}^N$ such that the intensity measurements $\{|\langle x, \varphi_n\rangle|^2\}_{n=1}^N$ enable successful recovery of the signal x (up to a global phase factor). In their seminal work on phase retrieval [8], Balan, Casazza and Edidin introduce the following property to analyze injectivity:

Definition 2.1. *An ensemble $\Phi = \{\varphi_n\}_{n=1}^N$ in $\mathbb{R}^M$ ($\mathbb{C}^M$) satisfies the complement property (CP) if for every $S \subseteq \{1, \ldots, N\}$, either $\{\varphi_n\}_{n \in S}$ or $\{\varphi_n\}_{n \in S^c}$ spans $\mathbb{R}^M$ ($\mathbb{C}^M$).*

Here and throughout, S^c denotes the set $\{1, \ldots, N\} \setminus S$. In the real case, the complement property is characteristic of injectivity, as demonstrated in [8]. The proof of this result is provided below; it contains several key insights which will be applied later.

Theorem 2.2. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ and the mapping $\mathcal{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n \rangle|^2$. Then $\mathcal{A}$ is injective if and only if Φ satisfies the complement property.*

Proof. We will prove both directions by obtaining the contrapositives.

($\Rightarrow$) Assume that Φ is not CP. Then there exists $S \subseteq \{1, \ldots, N\}$ such that neither $\{\varphi_n\}_{n \in S}$ nor $\{\varphi_n\}_{n \in S^c}$ spans $\mathbb{R}^M$. This implies that there are nonzero vectors $u, v \in \mathbb{R}^M$ such that $\langle u, \varphi_n \rangle = 0$ for all $n \in S$ and $\langle v, \varphi_n \rangle = 0$ for all $n \in S^c$. For each n, we then have

$$|\langle u \pm v, \varphi_n \rangle|^2 = |\langle u, \varphi_n \rangle|^2 \pm 2 \operatorname{Re}\langle u, \varphi_n \rangle \overline{\langle v, \varphi_n \rangle} + |\langle v, \varphi_n \rangle|^2 = |\langle u, \varphi_n \rangle|^2 + |\langle v, \varphi_n \rangle|^2.$$

Since $|\langle u + v, \varphi_n \rangle|^2 = |\langle u - v, \varphi_n \rangle|^2$ for every n, we have $\mathcal{A}(u + v) = \mathcal{A}(u - v)$. Moreover, u and v are nonzero by assumption, and so $u + v \neq \pm(u - v)$.

($\Leftarrow$) Assume that $\mathcal{A}$ is not injective. Then there exist vectors $x, y \in \mathbb{R}^M$ such that $x \neq \pm y$ and $\mathcal{A}(x) = \mathcal{A}(y)$. Taking $S := \{n : \langle x, \varphi_n \rangle = -\langle y, \varphi_n \rangle\}$, we have $\langle x + y, \varphi_n \rangle = 0$ for every $n \in S$. Otherwise when $n \in S^c$, we have $\langle x, \varphi_n \rangle = \langle y, \varphi_n \rangle$ and so $\langle x - y, \varphi_n \rangle = 0$. Furthermore, both $x + y$ and $x - y$ are nontrivial since $x \neq \pm y$, and so neither $\{\varphi_n\}_{n \in S}$ nor $\{\varphi_n\}_{n \in S^c}$ spans $\mathbb{R}^M$. □

Note that in [8] it is erroneously stated that the first part of the above proof also gives necessity of CP for injectivity in the complex case. Indeed, the proof demonstrates that $u+v \neq \pm(u-v)$, but fails to establish that $u+v \not\equiv (u-v) \bmod \mathbb{T}$; for instance, it could very well be the case that $u+v = \mathrm{i}(u-v)$, and so injectivity would not be violated *in the complex case*. A correct proof of the result in question is provided later (Theorem 2.6). In the meantime, we characterize injectivity in the complex case:

Theorem 2.3. *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ *and the mapping* $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ *defined by* $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$*. Viewing* $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ *as vectors in* $\mathbb{R}^{2M}$*, denote* $S(u) := \operatorname{span}_{\mathbb{R}}\{\varphi_n\varphi_n^* u\}_{n=1}^N$*. Then the following are equivalent:*

(a) $\mathcal{A}$ *is injective.*

(b) $\dim S(u) \geq 2M-1$ *for every* $u \in \mathbb{C}^M \setminus \{0\}$.

(c) $S(u) = \operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}^\perp$ *for every* $u \in \mathbb{C}^M \setminus \{0\}$.

Before proving this theorem, note that unlike the characterization in the real case, it is not clear whether this characterization can be tested in finite time; instead of being a statement about all (finitely many) partitions of $\{1,\ldots,N\}$, this is a statement about all $u \in \mathbb{C}^M \setminus \{0\}$. However, we can view this characterization as an analog to the real case in some sense: In the real case, the complement property is equivalent to having $\operatorname{span}\{\varphi_n\varphi_n^* u\}_{n=1}^N = \mathbb{R}^M$ for all $u \in \mathbb{R}^M \setminus \{0\}$. As the following proof makes precise, the fact that $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ fails to span all of $\mathbb{R}^{2M}$ is rooted in the fact that more information is lost with phase in the complex case.

Proof of Theorem 2.3. (a) ⇒ (c): Suppose $\mathcal{A}$ is injective. We need to show that $\{\varphi_n\varphi_n^* u\}_{n=1}^N$ spans the set of vectors orthogonal to $\mathrm{i}u$. Here, orthogonality is with respect to the real inner product, which can be expressed as $\langle a, b\rangle_{\mathbb{R}} = \operatorname{Re}\langle a, b\rangle$. Note that

$$|\langle u \pm v, \varphi_n\rangle|^2 = |\langle u, \varphi_n\rangle|^2 \pm 2\operatorname{Re}\langle u, \varphi_n\rangle\langle \varphi_n, v\rangle + |\langle v, \varphi_n\rangle|^2,$$

and so subtraction gives

$$|\langle u+v,\varphi_n\rangle|^2-|\langle u-v,\varphi_n\rangle|^2=4\operatorname{Re}\langle u,\varphi_n\rangle\langle\varphi_n,v\rangle=4\langle\varphi_n\varphi_n^*u,v\rangle_{\mathbb{R}}. \tag{1}$$

In particular, if the right-hand side of (1) is zero, then injectivity implies that there exists some ω of unit-modulus such that $u+v=\omega(u-v)$. Since $u\neq 0$, we know $\omega\neq -1$, and so rearranging gives

$$v=-\Big(\frac{1-\omega}{1+\omega}\Big)u=-\frac{(1-\omega)(1+\overline{\omega})}{|1+\omega|^2}u=\frac{2\operatorname{Im}\omega}{|1+\omega|^2}\,\mathrm{i}u.$$

This means $S(u)^\perp\subseteq\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}$. To prove $\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}\subseteq S(u)^\perp$, take $v=\alpha\mathrm{i}u$ for some $\alpha\in\mathbb{R}$ and define $\omega:=\frac{1+\alpha\mathrm{i}}{1-\alpha\mathrm{i}}$, which necessarily has unit-modulus. Then

$$u+v=u+\alpha\mathrm{i}u=(1+\alpha\mathrm{i})u=\frac{1+\alpha\mathrm{i}}{1-\alpha\mathrm{i}}(u-\alpha\mathrm{i}u)=\omega(u-v).$$

Thus, the left-hand side of (1) is zero, meaning $v\in S(u)^\perp$.

(b) ⇔ (c): First, (b) immediately follows from (c) since $\dim(\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\})=1$ for all $u\in\mathbb{C}^M\setminus\{0\}$. For the other direction, note that $\mathrm{i}u$ is necessarily orthogonal to every $\varphi_n\varphi_n^*u$:

$$\langle\varphi_n\varphi_n^*u,\mathrm{i}u\rangle_{\mathbb{R}}=\operatorname{Re}\langle\varphi_n\varphi_n^*u,\mathrm{i}u\rangle=\operatorname{Re}\langle u,\varphi_n\rangle\langle\varphi_n,\mathrm{i}u\rangle=-\operatorname{Re}\mathrm{i}|\langle u,\varphi_n\rangle|^2=0.$$

Thus, $\operatorname{span}_{\mathbb{R}}\{\mathrm{i}u\}\subseteq S(u)^\perp$ for all nonzero u. Since, by (b), $\dim S(u)^\perp\leq 1$, this then gives (c).

(c) ⇒ (a): This portion of the proof is inspired by Mukherjee's analysis in [80]. Suppose $\mathcal{A}(x)=\mathcal{A}(y)$. If $x=y$, we are done. Otherwise, $x-y\neq 0$, and so we may apply (c) to $u=x-y$. First, note that

$$\langle\varphi_n\varphi_n^*(x-y),x+y\rangle_{\mathbb{R}}=\operatorname{Re}\langle\varphi_n\varphi_n^*(x-y),x+y\rangle=\operatorname{Re}(x+y)^*\varphi_n\varphi_n^*(x-y),$$

and so expanding gives

$$\begin{aligned}\langle\varphi_n\varphi_n^*(x-y),x+y\rangle_\mathbb{R}&=\operatorname{Re}\Big(|\varphi_n^*x|^2-x^*\varphi_n\varphi_n^*y+y^*\varphi_n\varphi_n^*x-|\varphi_n^*y|^2\Big)\\&=\operatorname{Re}\Big(-x^*\varphi_n\varphi_n^*y+\overline{x^*\varphi_n\varphi_n^*y}\Big)=0.\end{aligned}$$

Since $x+y\in S(x-y)^\perp=\operatorname{span}_\mathbb{R}\{\mathrm{i}(x-y)\}$, there exists $\alpha\in\mathbb{R}$ such that $x+y=\alpha\mathrm{i}(x-y)$, and so rearranging gives $y=\frac{1-\alpha\mathrm{i}}{1+\alpha\mathrm{i}}x$, meaning $y\equiv x\bmod\mathbb{T}$. □

Theorem 2.3 leaves a lot to be desired; it is still unclear what it takes for a complex ensemble to yield injective intensity measurements. While in pursuit of a more clear understanding, we established the following bizarre characterization: A complex ensemble yields injective intensity measurements precisely when it yields injective *phase-only* measurements (in some sense). This is made more precise in the following theorem statement:

Theorem 2.4. *Consider $\Phi=\{\varphi_n\}_{n=1}^N\subseteq\mathbb{C}^M$ and the mapping $\mathcal{A}\colon\mathbb{C}^M/\mathbb{T}\rightarrow\mathbb{R}^N$ defined by $(\mathcal{A}(x))(n):=|\langle x,\varphi_n\rangle|^2$. Then $\mathcal{A}$ is injective if and only if the following statement holds: If for every $n=1,\ldots,N$, either $\arg(\langle x,\varphi_n\rangle^2)=\arg(\langle y,\varphi_n\rangle^2)$ or one of the sides is not well-defined, then $x=0$, $y=0$, or $y\equiv x\bmod\mathbb{R}\setminus\{0\}$.*

Proof. By Theorem 2.3, $\mathcal{A}$ is injective if and only if

$$\forall x\in\mathbb{C}^M\setminus\{0\},\qquad\operatorname{span}_\mathbb{R}\{\varphi_n\varphi_n^*x\}_{n=1}^N=\operatorname{span}_\mathbb{R}\{\mathrm{i}x\}^\perp.\tag{2}$$

Taking orthogonal complements of both sides, note that regardless of $x\in\mathbb{C}^M\setminus\{0\}$, we know $\operatorname{span}_\mathbb{R}\{\mathrm{i}x\}$ is necessarily a subset of $(\operatorname{span}_\mathbb{R}\{\varphi_n\varphi_n^*x\}_{n=1}^N)^\perp$, and so (2) is equivalent to

$$\begin{aligned}\forall x\in\mathbb{C}^M\setminus\{0\},\qquad&\operatorname{Re}\langle\varphi_n\varphi_n^*x,\mathrm{i}y\rangle=0\quad\forall n=1,\ldots,N\\&\qquad\Longrightarrow\quad y=0\text{ or }y\equiv x\bmod\mathbb{R}\setminus\{0\}.\end{aligned}$$

Thus, we need to determine when $\operatorname{Im}\langle x,\varphi_n\rangle\overline{\langle y,\varphi_n\rangle}=\operatorname{Re}\langle\varphi_n\varphi_n^*x,\mathrm{i}y\rangle=0$. We claim that this is true if and only if $\arg(\langle x,\varphi_n\rangle^2)=\arg(\langle y,\varphi_n\rangle^2)$ or one of the sides is not well-defined. To see this, we substitute $a:=\langle x,\varphi_n\rangle$ and $b:=\langle y,\varphi_n\rangle$. Then to complete the proof, it suffices to show that $\operatorname{Im}a\bar{b}=0$ if and only if $\arg(a^2)=\arg(b^2)$, $a=0$, or $b=0$.

($\Leftarrow$) If either a or b is zero, the result is immediate. Otherwise, if

$$2\arg(a)=\arg(a^2)=\arg(b^2)=2\arg(b),$$

then 2π divides $2(\arg(a)-\arg(b))$, and so $\arg(a\bar{b})=\arg(a)-\arg(b)$ is a multiple of π. This implies that $a\bar{b}\in\mathbb{R}$, and so $\operatorname{Im}a\bar{b}=0$.

($\Rightarrow$) Suppose $\operatorname{Im}a\bar{b}=0$. Taking the polar decompositions $a=re^{\mathrm{i}\theta}$ and $b=se^{\mathrm{i}\phi}$, we equivalently have that $rs\sin(\theta-\phi)=0$. Certainly, this can occur whenever r or s is zero, i.e., $a=0$ or $b=0$. Otherwise, a difference formula then gives $\sin\theta\cos\phi=\cos\theta\sin\phi$. From this, we know that if θ is an integer multiple of $\pi/2$, then ϕ is as well, and vice versa, in which case

$$\arg(a^2)=2\arg(a)=\pi=2\arg(b)=\arg(b^2).$$

Else, we can divide both sides by $\cos\theta\cos\phi$ to obtain $\tan\theta=\tan\phi$, from which it is evident that $\theta\equiv\phi \mod \pi$, and so $\arg(a^2)=2\arg(a)=2\arg(b)=\arg(b^2)$. □

This notion of injective phase-only measurements is similar to the idea of parallel rigidity in certain location estimation problems (for example, see [12] and references therein). It would be interesting to further investigate this relationship, although we will not do so here; at the very least, it is rather striking that injectivity is equivalent in both settings. We will actually use this result to (correctly) prove the necessity of CP for injectivity. First, we need the following lemma, which is interesting in its own right:

Lemma 2.5. *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ *and the mapping* $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ *defined by* $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$*. If* $\mathcal{A}$ *is injective, then the mapping* $\mathcal{B}\colon \mathbb{C}^M/\{\pm 1\} \to \mathbb{C}^N$ *defined by* $(\mathcal{B}(x))(n) := \langle x, \varphi_n\rangle^2$ *is also injective.*

Proof. Suppose $\mathcal{A}$ is injective. Then we have the following facts (one by definition, and the other by Theorem 2.4):

(i) If $|\langle x, \varphi_n\rangle|^2 = |\langle y, \varphi_n\rangle|^2$ for all $n = 1, \dots, N$, then $y \equiv x \bmod \mathbb{T}$.

(ii) If, for every $n \in \{1, \dots, N\}$, either $\arg(\langle x, \varphi_n\rangle^2) = \arg(\langle y, \varphi_n\rangle^2)$ or one of the sides is not well-defined, then $x = 0$, $y = 0$, or $y \equiv x \bmod \mathbb{R} \setminus \{0\}$.

Now suppose we have $\langle x, \varphi_n\rangle^2 = \langle y, \varphi_n\rangle^2$ for all $n = 1, \dots, N$. Then their moduli and arguments are also equal, and so (i) and (ii) both apply. Of course, $y \equiv x \bmod \mathbb{T}$ implies $x = 0$ if and only if $y = 0$. Otherwise both are nonzero, in which case there exists $\omega \in \mathbb{T} \cap \mathbb{R} \setminus \{0\} = \{\pm 1\}$ such that $y = \omega x$. In either case, $y \equiv x \bmod \{\pm 1\}$, so $\mathcal{B}$ is injective. □

Leveraging the injectivity of $\mathcal{B}$ modulo $\{\pm 1\}$, we may now extend the necessity of CP for injectivity to complex ensembles:

Theorem 2.6. *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ *and the mapping* $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ *defined by* $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$*. If* $\mathcal{A}$ *is injective, then* Φ *satisfies the complement property.*

Proof. Recall that if $\mathcal{A}$ is injective, then so is the mapping $\mathcal{B}$ of Lemma 2.5. Therefore, it suffices to show that Φ is CP if $\mathcal{B}$ is injective. To complete the proof, we will obtain the contrapositive (note the similarity to the proof of Theorem 2.2). Suppose Φ is not CP. Then there exists $S \subseteq \{1, \dots, N\}$ such that neither $\{\varphi_n\}_{n\in S}$ nor $\{\varphi_n\}_{n\in S^{\mathrm{c}}}$ spans $\mathbb{C}^M$. This implies that there are nonzero vectors $u, v \in \mathbb{C}^M$ such that $\langle u, \varphi_n\rangle = 0$ for all $n \in S$ and $\langle v, \varphi_n\rangle = 0$ for all $n \in S^{\mathrm{c}}$. For each n, we then have

$$\langle u \pm v, \varphi_n\rangle^2 = \langle u, \varphi_n\rangle^2 \pm 2\langle u, \varphi_n\rangle\langle v, \varphi_n\rangle + \langle v, \varphi_n\rangle^2 = \langle u, \varphi_n\rangle^2 + \langle v, \varphi_n\rangle^2.$$

Since $\langle u+v,\varphi_n\rangle^2=\langle u-v,\varphi_n\rangle^2$ for every n, we have $\mathcal{B}(u+v)=\mathcal{B}(u-v)$. Moreover, u and v are nonzero by assumption, and so $u+v\neq\pm(u-v)$. □

Note that the complement property is necessary but not sufficient for injectivity. To see this, consider the measurement vectors $(1,0)$, $(0,1)$ and $(1,1)$ in $\mathbb{C}^2$. These certainly satisfy the complement property, but $\mathcal{A}((1,\mathrm{i}))=(1,1,2)=\mathcal{A}((1,-\mathrm{i}))$, despite the fact that $(1,\mathrm{i})\not\equiv(1,-\mathrm{i})\bmod\mathbb{T}$; in general, real measurement vectors fail to yield injective intensity measurements in the complex setting since they do not distinguish complex conjugates. Indeed, we have yet to find a "good" sufficient condition for injectivity in the complex case. As an analogy for what we really want, consider the notion of *full spark*: An ensemble $\{\varphi_n\}_{n=1}^N\subseteq\mathbb{R}^M$ is said to be full spark if every subcollection of M vectors spans $\mathbb{R}^M$. It is easy to see that full spark ensembles with $N\geq 2M-1$ necessarily satisfy the complement property (thereby implying injectivity in the real case), since in this case

$$\min_{S\subseteq\{1,2,\ldots,N\}}\Big\{\max\{|S|,|S^\mathrm{c}|\}\Big\}=M,$$

and so it is guaranteed that one of the sets $\{\varphi_n\}_{n\in S}$ or $\{\varphi_n\}_{n\in S^\mathrm{c}}$ spans. Furthermore, the notion of full spark is simple enough to admit deterministic constructions [4,81]. Deterministic measurement ensembles are particularly desirable for the complex case, and so finding a good sufficient condition for injectivity is an important problem that remains open.

2.2 Towards a rank-nullity theorem for phase retrieval

If one thinks of a matrix Φ as being built one column at a time, then the rank-nullity theorem states that each column contributes to either the column space or the null space. If the columns are then used as linear measurement vectors (say we take measurements $y=\Phi^*x$ of a vector x), then the column space of Φ gives the subspace that is actually sampled, and the null space captures the algebraic nature of the measurements' redundancy. Therefore, an efficient sampling of an entire vector space would apply a matrix Φ with a small null space and large column space (e.g., an invertible square matrix). How do we find such a sampling with intensity measurements? The following makes this question more precise:

Problem 2.7. *For any dimension M, what is the smallest number $N^*(M)$ of injective intensity measurements, and how do we design such measurement vectors?*

To be clear, this problem was completely solved in the real case by Balan, Casazza and Edidin [8]. Indeed, Theorem 2.2 immediately implies that $2M-2$ intensity measurements are necessarily not injective, and furthermore that $2M-1$ measurements are injective if and only if the measurement vectors are full spark. As such, we will focus our attention to the complex case.

In the complex case, Problem 2.7 has some history in the quantum mechanics literature. For example, [87] presents *Wright's conjecture* that three observables suffice to uniquely determine any pure state. In phase retrieval parlance, the conjecture states that there exist unitary matrices U_1, U_2 and U_3 such that $\Phi = [U_1\ U_2\ U_3]$ yields injective intensity measurements. Note that Wright's conjecture actually implies that $N^*(M) \leq 3M-2$; indeed, U_1 determines the norm (squared) of the signal, rendering the last column of both U_2 and U_3 unnecessary. Finkelstein [56] later proved that $N^*(M) \geq 3M-2$; combined with Wright's conjecture, this led many to believe that $N^*(M) = 3M-2$ (for example, see [24]). However, both this and Wright's conjecture were recently disproved in [62], in which Heinosaari, Mazzarella and Wolf invoked embedding theorems from differential geometry to prove that

$$N^*(M) \geq \begin{cases} 4M-2\alpha(M-1)-3 & \text{for all } M \\ 4M-2\alpha(M-1)-2 & \text{if } M \text{ is odd and } \alpha(M-1) = 2 \bmod 4 \\ 4M-2\alpha(M-1)-1 & \text{if } M \text{ is odd and } \alpha(M-1) = 3 \bmod 4, \end{cases} \tag{3}$$

where $\alpha(M-1) \leq \log_2(M)$ is the number of 1's in the binary representation of $M-1$. By comparison, Balan, Casazza and Edidin [8] proved that $N^*(M) \leq 4M-2$, and so we at least have the asymptotic expression $N^*(M) = (4+o(1))M$.

At this point, we should clarify some intuition for $N^*(M)$ by explaining the nature of these best known lower and upper bounds. First, the lower bound (3) follows from an older result that complex projective space $\mathbb{C}\mathbf{P}^n$ does not smoothly embed into $\mathbb{R}^{4n-2\alpha(n)}$ (and other slight refinements which depend on n); this is due to Mayer [75], but we highly recommend James's survey on the topic [66]. To prove (3) from this, suppose $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \rightarrow \mathbb{R}^N$ were injective. Then $\mathcal{E}$ defined by $\mathcal{E}(x) := \mathcal{A}(x)/\|x\|^2$ embeds $\mathbb{C}\mathbf{P}^{M-1}$ into $\mathbb{R}^N$, and as Heinosaari et al. show, the embedding is necessarily smooth; considering $\mathcal{A}(x)$ is made up of rather simple polynomials, the fact that $\mathcal{E}$ is smooth should not come as a surprise. As such, the nonembedding result produces the best known lower bound. To evaluate this bound, first note that Milgram [77] constructs an embedding of $\mathbb{C}\mathbf{P}^n$ into $\mathbb{R}^{4n-\alpha(n)+1}$, establishing the importance of the $\alpha(n)$ term, but the constructed embedding does not correspond to an intensity measurement process. In order to relate these embedding results to our problem, consider the real case: It is known that for odd $n \geq 7$, real projective space $\mathbb{R}\mathbf{P}^n$ smoothly embeds into $\mathbb{R}^{2n-\alpha(n)+1}$ [86], which means the analogous lower bound for the real case would necessarily be smaller than

$$2(M-1)-\alpha(M-1)+1 = 2M-\alpha(M-1)-1 < 2M-1.$$

This indicates that the $\alpha(M-1)$ term in (3) might be an artifact of the proof technique, rather than of $N^*(M)$.

There is also some intuition to be gained from the upper bound $N^*(M) \leq 4M-2$, which Balan et al. [8] proved by applying certain techniques from algebraic geometry (some of which will be applied later in this section). In fact, their result actually gives that $4M-2$ or more measurement vectors, if chosen *generically*, will yield injective intensity measurements; here, generic is a technical term involving the Zariski topology, but it can be thought of as some undisclosed property which is satisfied with probability 1 by measurement vectors drawn from continuous distributions. This leads us to think that $N^*(M)$ generic measurement vectors might also yield injectivity.

The lemma that follows will help to refine our intuition for $N^*(M)$, and it will also play a key role in the main theorems of this section (a similar result appears in [62]). Before stating the result, define the real M^2-dimensional space $\mathbb{H}^{M\times M}$ of self-adjoint $M \times M$ matrices; note that this is not a vector space over the complex numbers since the diagonal of a self-adjoint matrix must be real. Given an ensemble of measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$, define the *super analysis operator* $\mathbf{A}\colon \mathbb{H}^{M\times M} \rightarrow \mathbb{R}^N$ by $(\mathbf{A}H)(n) = \langle H, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}}$; here, $\langle\cdot,\cdot\rangle_{\mathrm{HS}}$ denotes the Hilbert-Schmidt inner product, which induces the Frobenius matrix norm. Note that $\mathbf{A}$ is a linear operator, and yet

$$\begin{aligned}(\mathbf{A}xx^*)(n) &= \langle xx^*, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}} = \mathrm{Tr}[\varphi_n\varphi_n^* xx^*]\\ &= \mathrm{Tr}[\varphi_n^* xx^*\varphi_n] = \varphi_n^* xx^* \varphi_n = |\langle x, \varphi_n\rangle|^2 = (\mathcal{A}(x))(n).\end{aligned}$$

In words, the class of vectors identified with x modulo $\mathbb{T}$ can be “lifted” to xx^*, thereby linearizing the intensity measurement process at the price of squaring the dimension of the vector space of interest; this identification has been exploited by some of the most noteworthy strides in modern phase retrieval [7,27]. As the following lemma shows, this identification can also be used to characterize injectivity:

Lemma 2.8. *$\mathcal{A}$ is not injective if and only if there exists a matrix of rank* 1 *or* 2 *in the null space of* $\mathbf{A}$.

Proof. ($\Rightarrow$) If $\mathcal{A}$ is not injective, then there exist $x, y \in \mathbb{C}^M/\mathbb{T}$ with $x \not\equiv y \bmod \mathbb{T}$ such that $\mathcal{A}(x) = \mathcal{A}(y)$. That is, $\mathbf{A}xx^* = \mathbf{A}yy^*$, and so $xx^* - yy^*$ is in the null space of $\mathbf{A}$.

($\Leftarrow$) First, suppose there is a rank-1 matrix H in the null space of $\mathbf{A}$. Then there exists $x \in \mathbb{C}^M$ such that $H = xx^*$ and

$$(\mathcal{A}(x))(n) = (\mathbf{A}xx^*)(n) = 0 = (\mathcal{A}(0))(n).$$

But $x \not\equiv 0 \bmod \mathbb{T}$, and so $\mathcal{A}$ is not injective. Now suppose there is a rank-2 matrix H in the null space of $\mathbf{A}$. Then by the spectral theorem, there are orthonormal $u_1, u_2 \in \mathbb{C}^M$ and nonzero $\lambda_1 \geq \lambda_2$ such that $H = \lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^*$. Since H is in the null space of $\mathbf{A}$, the following holds for every n:

$$0 = \langle H, \varphi_n \varphi_n^* \rangle_{\mathrm{HS}} = \langle \lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^*, \varphi_n \varphi_n^* \rangle_{\mathrm{HS}} = \lambda_1 |\langle u_1, \varphi_n \rangle|^2 + \lambda_2 |\langle u_2, \varphi_n \rangle|^2. \quad (4)$$

Taking $x := |\lambda_1|^{1/2} u_1$ and $y := |\lambda_2|^{1/2} u_2$, note that $y \not\equiv x \bmod \mathbb{T}$ since they are nonzero and orthogonal. We claim that $\mathcal{A}(x) = \mathcal{A}(y)$, which would complete the proof. If λ_1 and λ_2 have the same sign, then by (4), $|\langle x, \varphi_n \rangle|^2 + |\langle y, \varphi_n \rangle|^2 = 0$ for every n, meaning $|\langle x, \varphi_n \rangle|^2 = 0 = |\langle y, \varphi_n \rangle|^2$. Otherwise, $\lambda_1 > 0 > \lambda_2$, and so

$$xx^* - yy^* = \lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^* = A$$

is in the null space of $\mathbf{A}$, meaning $\mathcal{A}(x) = \mathbf{A}xx^* = \mathbf{A}yy^* = \mathcal{A}(y)$. □

Lemma 2.8 indicates that we want the null space of $\mathbf{A}$ to avoid nonzero matrices of rank ≤ 2. Intuitively, this is easier when the "dimension" of this set of matrices is small. To get some idea of this dimension, count real degrees of freedom: By the spectral theorem, almost every matrix in $\mathbb{H}^{M \times M}$ of rank ≤ 2 can be uniquely expressed as $\lambda_1 u_1 u_1^* + \lambda_2 u_2 u_2^*$ with $\lambda_1 \leq \lambda_2$. Here, (λ_1, λ_2) has two degrees of freedom. Next, u_1 can be any vector in $\mathbb{C}^M$, except its norm must be 1. Also, since u_1 is only unique up to global phase, we take its first entry to be nonnegative without loss of generality. Given the norm and phase constraints, u_1 has a total of $2M - 2$ real degrees of freedom. Finally, u_2 has the same norm and phase constraints, but it must also be orthogonal to u_1, that is, $\mathrm{Re}\langle u_2, u_1\rangle = \mathrm{Im}\langle u_2, u_1\rangle = 0$. As such, u_2 has $2M-4$ real degrees of freedom. All together, we can expect the set of matrices in question to have $2+(2M-2)+(2M-4) = 4M-4$ real dimensions.

If the set S of matrices of rank ≤ 2 formed a subspace of $\mathbb{H}^{M\times M}$ (it doesn't), then we could expect the null space of $\mathbf{A}$ to intersect that subspace nontrivially whenever $\dim\mathrm{null}(\mathbf{A})+(4M-4) > \dim(\mathbb{H}^{M\times M}) = M^2$. By the rank-nullity theorem, this would indicate that injectivity requires

$$N \geq \mathrm{rank}(\mathbf{A}) = M^2 - \dim\mathrm{null}(\mathbf{A}) \geq 4M-4. \tag{5}$$

Of course, this logic is not technically valid since S is not a subspace. It is, however, a special kind of set: a real projective variety. To see this, we first show that it is a real algebraic variety, specifically, the set of members of $\mathbb{H}^{M\times M}$ for which all 3×3 minors are zero. Of course, by the rank constraint, every member of S has this minor property. Next, we show that members of S are the only matrices with this property: If the rank of a given matrix is ≥ 3, then it has an $M\times 3$ submatrix of linearly independent columns, and since the rank of its transpose is also ≥ 3, this $M\times 3$ submatrix must have 3 linearly independent rows, thereby implicating a full-rank 3×3 submatrix. This variety is said to be projective because it is closed under scalar multiplication. If S were a projective variety over an algebraically closed field (it's not), then the projective dimension theorem (Theorem 7.2 of [61]) says that S intersects $\mathrm{null}(\mathbf{A})$ nontrivially whenever the dimensions are large enough: $\dim\mathrm{null}(\mathbf{A}) + \dim S > \dim\mathbb{H}^{M\times M}$, thereby implying that injectivity requires (5). Unfortunately, this theorem is not valid when the field is $\mathbb{R}$; for example, the cone defined by $x^2+y^2-z^2=0$ in $\mathbb{R}^3$ is a projective variety of dimension 2, but its intersection with the 2-dimensional xy-plane is trivial, despite the fact that $2+2>3$.

In the absence of a proof, we pose the natural conjecture:

Conjecture 2.9 (The $4M-4$ Conjecture)**.** *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^M$ and the mapping $\mathcal{A}\colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. If $M \geq 2$, then the following statements hold:*

(a) If $N < 4M-4$, then $\mathcal{A}$ is not injective.

(b) If $N \geq 4M-4$, then $\mathcal{A}$ is injective for generic Φ.

For the sake of clarity, we now explicitly state what is meant by the word "generic." As indicated above, a real algebraic variety is the set of common zeros of a finite set of polynomials with real coefficients. Taking all such varieties in $\mathbb{R}^n$ to be closed sets defines the *Zariski topology* on $\mathbb{R}^n$. Viewing Φ as a member of $\mathbb{R}^{2MN}$, we then say a *generic* Φ is any member of some undisclosed nonempty Zariski-open subset of $\mathbb{R}^{2MN}$. Considering Zariski-open sets are either empty or dense with full measure, genericity is a particularly strong property. As such, another way to state part (b) of the $4M-4$ conjecture is "If $N \geq 4M-4$, then there exists a real algebraic variety $V \subseteq \mathbb{R}^{2MN}$ such that $\mathcal{A}$ is injective for every $\Phi \notin V$." Note that the work of Balan, Casazza and Edidin [8] already proves this for $N \geq 4M-2$, and in the time since we initially posed this conjecture [10, 11], Conca, Edidin, Hering and Vinzant [38] proved it for the case $M = 2^m + 1$, where m is any positive integer. Furthermore, Conca et al. successfully established part (b) by using techniques from algebraic geometry to show that the set of non-injective ensembles is a subset of a proper real algebraic variety and, hence, a Zariski closed set [38].

At this point, it is fitting to mention that after initially formulating this conjecture, Bodmann presented a Vandermonde construction of $4M-4$ injective intensity measurements at a phase retrieval workshop at the Erwin Schrödinger International Institute for Mathematical Physics. The result has since been documented in [17], and it establishes one consequence of the $4M-4$ conjecture: $N^*(M) \leq 4M-4$.

As incremental progress toward solving the $4M-4$ conjecture, we have the following result:

Theorem 2.10. *The $4M-4$ Conjecture is true when $M=2$.*

Proof. (a) Since $\mathbf{A}$ is a linear map from 4-dimensional real space to N-dimensional real space, the null space of $\mathbf{A}$ is necessarily nontrivial by the rank-nullity theorem. Furthermore, every nonzero member of this null space has rank 1 or 2, and so Lemma 2.8 gives that $\mathcal{A}$ is not injective.

(b) Consider the following matrix formed by 16 real variables:

$$\Phi(x)=\begin{bmatrix} x_1+\mathrm{i}x_2 & x_5+\mathrm{i}x_6 & x_9+\mathrm{i}x_{10} & x_{13}+\mathrm{i}x_{14} \\ x_3+\mathrm{i}x_4 & x_7+\mathrm{i}x_8 & x_{11}+\mathrm{i}x_{12} & x_{15}+\mathrm{i}x_{16} \end{bmatrix}. \tag{6}$$

If we denote the nth column of $\Phi(x)$ by $\varphi_n(x)$, then we have that $\mathcal{A}$ is injective precisely when $x\in\mathbb{R}^{16}$ produces a basis $\{\varphi_n(x)\varphi_n(x)^*\}_{n=1}^4$ for the space of 2×2 self-adjoint operators. Indeed, in this case zz^* is uniquely determined by $\mathbf{A}zz^*=\{\langle zz^*,\varphi_n(x)\varphi_n(x)^*\rangle_{\mathrm{HS}}\}_{n=1}^4=\mathcal{A}(z)$, which in turn determines z up to a global phase factor. Let $\mathbf{A}(x)$ be the 4×4 matrix representation of the super analysis operator, whose nth row gives the coordinates of $\varphi_n(x)\varphi_n(x)^*$ in terms of some basis for $\mathbb{H}^{2\times 2}$, say

$$\left\{\begin{bmatrix}1&0\\0&1\end{bmatrix},\begin{bmatrix}0&0\\0&1\end{bmatrix},\frac{1}{\sqrt{2}}\begin{bmatrix}0&1\\1&0\end{bmatrix},\frac{1}{\sqrt{2}}\begin{bmatrix}0&\mathrm{i}\\-\mathrm{i}&0\end{bmatrix}\right\}. \tag{7}$$

Then $V=\{x:\operatorname{Re}\det\mathbf{A}(x)=\operatorname{Im}\det\mathbf{A}(x)=0\}$ is a real algebraic variety in $\mathbb{R}^{16}$, and we see that $\mathcal{A}$ is injective whenever $x\in V^{\mathrm{c}}$. Since V^{c} is Zariski-open, it is either empty or dense with full measure. In fact, V^{c} is not empty, since we may take x such that

$$\Phi(x)=\begin{bmatrix}1&0&1&1\\0&1&1&\mathrm{i}\end{bmatrix},$$

as indicated in Theorem 4.1 of [6]. Therefore, V^{c} is dense with full measure. □

Algorithm 1 The HMW test for injectivity when $M = 3$

```
Input: Measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{C}^3$
Output: Whether $\mathcal{A}$ is injective
  Define $\mathbf{A}\colon \mathbb{H}^{3\times 3} \to \mathbb{R}^N$ such that $\mathbf{A}H = \{\langle H, \varphi_n\varphi_n^*\rangle_{\mathrm{HS}}\}_{n=1}^N$
  if $\dim\operatorname{null}(\mathbf{A}) = 0$ then
    Output: "INJECTIVE"                {if $\mathbf{A}$ is injective, then $\mathcal{A}$ is injective}
  else
    Pick $H \in \operatorname{null}(\mathbf{A})$, $H \neq 0$
    if $\dim\operatorname{null}(\mathbf{A}) = 1$ and $\det(H) \neq 0$ then
      Output: "INJECTIVE"              {if $\mathbf{A}$ only maps nonsingular matrices
                                        to zero, then $\mathcal{A}$ is injective}
    else
      Output: "NOT INJECTIVE"          {in the remaining case, $\mathbf{A}$ maps differences
                                        of rank-1 matrices to zero}
    end if
  end if
```

We also have a proof for the $M = 3$ case, but we first introduce Algorithm 1, namely the *HMW test* for injectivity; we name it after Heinosaari, Mazarella and Wolf, who implicitly introduce this algorithm in their paper [62].

Theorem 2.11 (cf. Proposition 6 in [62])**.** *When $M = 3$, the HMW test correctly determines whether $\mathcal{A}$ is injective.*

Proof. First, if $\mathbf{A}$ is injective, then $\mathcal{A}(x) = \mathbf{A}xx^* = \mathbf{A}yy^* = \mathcal{A}(y)$ if and only if $xx^* = yy^*$, i.e., $y \equiv x \bmod \mathbb{T}$. Next, suppose $\mathbf{A}$ has a 1-dimensional null space. Then Lemma 2.8 gives that $\mathcal{A}$ is injective if and only if the null space of $\mathbf{A}$ is spanned by a matrix of full rank. Finally, if the dimension of the null space is 2 or more, then there exist linearly independent (nonzero) matrices A and B in this null space. If $\det(A) = 0$, then it must have rank 1 or 2, and so Lemma 2.8 gives that $\mathcal{A}$ is not injective. Otherwise, consider the map

$$f\colon t \mapsto \det(A\cos t + B\sin t) \qquad \forall t \in [0,\pi].$$

Since $f(0) = \det(A)$ and $f(\pi) = \det(-A) = (-1)^3\det(A) = -\det(A)$, the intermediate value theorem gives that there exists $t_0 \in [0,\pi]$ such that $f(t_0) = 0$, i.e., the matrix $A\cos t_0 + B\sin t_0$ is singular. Moreover, this matrix is nonzero since A and B are linearly independent, and so its rank is either 1 or 2. Lemma 2.8 then gives that $\mathcal{A}$ is not injective. □

As an example, we may run the HMW test on the columns of the following matrix:

$$\Phi = \begin{bmatrix} 2 & 1 & 1 & 0 & 0 & 0 & 1 & \mathrm{i} \\ -1 & 0 & 0 & 1 & 1 & -1 & -2 & 2 \\ 0 & 1 & -1 & 1 & -1 & 2\mathrm{i} & \mathrm{i} & -1 \end{bmatrix}. \tag{8}$$

In this case, the null space of $\mathbf{A}$ is 1-dimensional and spanned by a nonsingular matrix. As such, $\mathcal{A}$ is injective. We will see that the HMW test has a few important applications. First, we use it to prove the $4M-4$ Conjecture in the $M=3$ case:

Theorem 2.12. *The $4M-4$ Conjecture is true when $M=3$.*

Proof. (a) Suppose $N < 4M-4 = 8$. Then by the rank-nullity theorem, the super analysis operator $\mathbf{A}\colon \mathbb{H}^{3\times 3} \to \mathbb{R}^N$ has a null space of at least 2 dimensions, and so by the HMW test, $\mathcal{A}$ is not injective.

(b) Consider a 3×8 matrix of real variables $\Phi(x)$ similar to (6). Then $\mathcal{A}$ is injective whenever $x \in \mathbb{R}^{48}$ produces an ensemble $\{\varphi_n(x)\}_{n=1}^{8} \subseteq \mathbb{C}^3$ that passes the HMW test. To pass, the rank-nullity theorem says that the null space of the super analysis operator must be 1-dimensional and spanned by a nonsingular matrix. We use an orthonormal basis for $\mathbb{H}^{3\times 3}$ similar to (7) to find an 8×9 matrix representation of the super analysis operator $\mathbf{A}(x)$; it is easy to check that the entries of this matrix (call it $\mathbf{A}(x)$) are polynomial functions of x. Consider the matrix

$$B(x,y) = \begin{bmatrix} y^{\mathrm{T}} \\ \mathbf{A}(x) \end{bmatrix},$$

and let $u(x)$ denote the vector of $(1,j)$th cofactors of $B(x,y)$. Then $\langle y, u(x)\rangle = \det(B(x,y))$. This implies that $u(x)$ is in the null space of $\mathbf{A}(x)$, since each row of $\mathbf{A}(x)$ is necessarily orthogonal to $u(x)$.

We claim that $u(x) = 0$ if and only if the dimension of the null space of $\mathbf{A}(x)$ is 2 or more, that is, the rows of $\mathbf{A}(x)$ are linearly dependent. First, ($\Leftarrow$) is true since the entries of $u(x)$ are signed determinants of 8×8 submatrices of $\mathbf{A}(x)$, which are necessarily zero by the linear dependence of the rows. For ($\Rightarrow$), we have that $0 = \langle y, 0\rangle = \langle y, u(x)\rangle = \det(B(x,y))$ for all $y \in \mathbb{R}^9$. That is, even if y is nonzero and orthogonal to the rows of $\mathbf{A}(x)$, the rows of $B(x,y)$ are linearly dependent, and so the rows of $\mathbf{A}(x)$ must be linearly dependent. This proves the intermediate claim.

We now use the claim to prove the result. The entries of $u(x)$ are coordinates of a matrix $U(x) \in \mathbb{H}^{3\times 3}$ in the same basis as before. Note that the entries of $U(x)$ are polynomials of x. Furthermore, $\mathcal{A}$ is injective if and only if $\det U(x) \neq 0$. To see this, observe three cases:

Case I: $U(x) = 0$, i.e., $u(x) = 0$, or equivalently, $\dim \operatorname{null}(\mathbf{A}(x)) \geq 2$. By the HMW test, $\mathcal{A}$ is not injective.

Case II: The null space is spanned by $U(x) \neq 0$, but $\det U(x) = 0$. By the HMW test, $\mathcal{A}$ is not injective.

Case III: The null space is spanned by $U(x) \neq 0$, and $\det U(x) \neq 0$. By the HMW test, $\mathcal{A}$ is injective.

Defining the real algebraic variety $V = \{x : \det U(x) = 0\} \subseteq \mathbb{R}^{48}$, we then have that $\mathcal{A}$ is injective precisely when $x \in V^{\mathrm{c}}$. Since V^{c} is Zariski-open, it is either empty or dense with full measure, but it is nonempty since (8) passes the HMW test. Therefore, V^{c} is dense with full measure. □

To be clear, this result has since been proven as part of a larger class of ensembles for which the conjecture holds, namely, the case $M = 2^m + 1$ for any positive integer m [38]. In fact, Conca, Edidin, Hering and Vinzant prove much more:

Theorem 2.13 (Theorem 1.1 and Proposition 5.4 in [38]).

(a) If m is any positive integer, then part (a) of the $4M-4$ Conjecture is true for $M = 2^m + 1$.

(b) Part (b) of the $4M-4$ Conjecture is true.

As a consequence of Theorems 2.10 and 2.13, the first remaining open case of the $4M-4$ Conjecture is $M = 4$.

Now recall Wright's conjecture: there exist unitary matrices U_1, U_2 and U_3 such that $\Phi = [U_1\ U_2\ U_3]$ yields injective intensity measurements. Also recall that Wright's conjecture implies $N^*(M) \leq 3M-2$. Again, both of these were disproved by Heinosaari et al. [62] using deep results in differential geometry. Alternatively, Theorem 2.12 also disproves these in the case where $M = 3$, since $N^*(3) = 4(3)-3 = 8 > 7 = 3(3)-2$.

Note that the HMW test can be used to test for injectivity in three dimensions regardless of the number of measurement vectors. As such, it can be used to evaluate ensembles of 3×3 unitary matrices for quantum mechanics. For example, consider the 3×3 fractional discrete Fourier transform, defined in [22] using discrete Hermite-Gaussian functions:

$$F^\alpha = \frac{1}{6}\begin{bmatrix} 3+\sqrt{3} & \sqrt{3} & \sqrt{3} \\ \sqrt{3} & \frac{3-\sqrt{3}}{2} & \frac{3-\sqrt{3}}{2} \\ \sqrt{3} & \frac{3-\sqrt{3}}{2} & \frac{3-\sqrt{3}}{2} \end{bmatrix} + \frac{e^{\alpha i\pi}}{6}\begin{bmatrix} 3-\sqrt{3} & -\sqrt{3} & -\sqrt{3} \\ -\sqrt{3} & \frac{3+\sqrt{3}}{2} & \frac{3+\sqrt{3}}{2} \\ -\sqrt{3} & \frac{3+\sqrt{3}}{2} & \frac{3+\sqrt{3}}{2} \end{bmatrix} + \frac{e^{\alpha i\pi/2}}{2}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

It can be shown by the HMW test that $\Phi = [I\ F^{1/2}\ F\ F^{3/2}]$ yields injective intensity measurements. This leads to the following refinement of Wright's conjecture:

Conjecture 2.14. *Let F denote the $M \times M$ discrete fractional Fourier transform defined in [22]. Then for every $M \geq 3$, $\Phi = [I\ F^{1/2}\ F\ F^{3/2}]$ yields injective intensity measurements.*

This conjecture can be viewed as the discrete analog to the work of Jaming [67], in which ensembles of *continuous* fractional Fourier transforms are evaluated for injectivity.

2.3 Achieving injectivity with $4M - 4$ *intensity measurements*

In this section, we provide an ensemble of $4M - 4$ measurement vectors which yield injective intensity measurements for $\mathbb{C}^M$. The vectors in this ensemble are modulated discrete cosine functions, and they are explicitly constructed at the end of this section. We start here by motivating the construction, specifically by identifying the significance of *circular autocorrelation*, which we define in (9) below.

Consider the P-dimensional complex vector space

$$\ell(\mathbb{Z}_P) := \{u\colon \mathbb{Z} \to \mathbb{C} : u(p+P) = u(p),\ \forall p \in \mathbb{Z}\}.$$

The discrete Fourier basis in $\ell(\mathbb{Z}_P)$ is the sequence of P vectors $\{f_q\}_{q\in\mathbb{Z}_P}$ defined by $f_q(p) := e^{2\pi i pq/P}$ (the notation "$q \in \mathbb{Z}_P$" is taken to mean a set of coset representatives of $\mathbb{Z}$ with respect to the subgroup $P\mathbb{Z}$). The discrete Fourier transform (DFT) on $\mathbb{Z}_P$ is $F^*\colon \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$, with corresponding inverse DFT $(F^*)^{-1} = \frac{1}{P}F$, defined by

$$(F^*u)(q) = \langle u, f_q\rangle = \sum_{p\in\mathbb{Z}_P} u(p)e^{-2\pi i pq/P},$$
$$(Fv)(p) = \sum_{q\in\mathbb{Z}_P} v(q)f_q(p) = \sum_{q\in\mathbb{Z}_P} v(q)e^{2\pi i pq/P}.$$

Now let $T^p\colon \ell(\mathbb{Z}_P)\to\ell(\mathbb{Z}_P)$ be the translation operator $(T^pu)(p'):=u(p'-p)$. The circular autocorrelation of u is then $\operatorname{CirAut}(u)\in\ell(\mathbb{Z}_P)$, defined entrywise by

$$(\operatorname{CirAut}(u))(p):=\langle u,T^pu\rangle=\sum_{p'\in\mathbb{Z}_P}u(p')\overline{u(p'-p)}. \tag{9}$$

Consider the DFT of a circular autocorrelation:

$$\begin{aligned}(F^*\operatorname{CirAut}(u))(q)&=\sum_{p\in\mathbb{Z}_P}\sum_{p'\in\mathbb{Z}_P}u(p')\overline{u(p'-p)}e^{-2\pi ipq/P}\\&=\sum_{p'\in\mathbb{Z}_P}u(p')e^{-2\pi ip'q/P}\overline{\bigg(\sum_{p\in\mathbb{Z}_P}u(p'-p)e^{-2\pi i(p'-p)q/P}\bigg)}\\&=\sum_{p'\in\mathbb{Z}_P}u(p')e^{-2\pi ip'q/P}\overline{\bigg(\sum_{p''\in\mathbb{Z}_P}u(p'')e^{-2\pi ip''q/P}\bigg)}=|\langle u,f_q\rangle|^2.\end{aligned} \tag{10}$$

As such, if one has the intensity measurements $\{|\langle u,f_q\rangle|^2\}_{q\in\mathbb{Z}_P}$, then one may compute the circular autocorrelation $\operatorname{CirAut}(u)$ by applying the inverse DFT. In order to perform phase retrieval from $\{|\langle u,f_q\rangle|^2\}_{q\in\mathbb{Z}_P}$, it therefore suffices to determine u from $\operatorname{CirAut}(u)$. This is the motivation for the approach in this section.

To see how to "invert" CirAut, let's consider an example. Take $x=(a,b,c)\in\mathbb{C}^3$ and consider the circular autocorrelation of x as a signal in $\ell(\mathbb{Z}_3)$:

$$\operatorname{CirAut}(x)=(|a|^2+|b|^2+|c|^2,a\overline{c}+b\overline{a}+c\overline{b},a\overline{b}+b\overline{c}+c\overline{a}).$$

Notice that every entry of $\operatorname{CirAut}(x)$ is a nonlinear combination of the entries of x, from which it is unclear how to compute the entries of x. To simplify the structure, we pad x with zeros and enforce even symmetry; then the circular autocorrelation of $u:=(2a,b,c,0,0,0,0,c,b)\in\ell(\mathbb{Z}_9)$ is

$$\begin{aligned}\operatorname{CirAut}(u)=(4|a|^2+|b|^2+|c|^2,2\operatorname{Re}(2a\overline{b}+b\overline{c}),|b|^2+4\operatorname{Re}(a\overline{c}),2\operatorname{Re}(b\overline{c}),|c|^2,\\|c|^2,2\operatorname{Re}(b\overline{c}),|b|^2+4\operatorname{Re}(a\overline{c}),2\operatorname{Re}(2a\overline{b}+b\overline{c})).\end{aligned} \tag{11}$$

Although it still appears rather complicated, this circular autocorrelation actually lends itself well to recovering the entries of x.

Before explaining this further, first note that $9 = 4(3) - 3$, and we can generalize our mapping $x \mapsto u$ by sending vectors in $\mathbb{C}^M$ to members of $\ell(\mathbb{Z}_{4M-3})$. To make this clear, consider the reversal operator $R\colon \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ defined by $(Ru)(p) = u(-p)$. Then given a vector $x \in \mathbb{C}^M$, padding with zeros and enforcing even symmetry is equivalent to embedding x in $\ell(\mathbb{Z}_{4M-3})$ by appending $3M - 3$ zeros to x and then taking $u = x + Rx \in \ell(\mathbb{Z}_{4M-3})$. (From this point forward, "x" is used to represent both the original signal in $\mathbb{C}^M$ and the version of x embedded in $\ell(\mathbb{Z}_{4M-3})$ via zero-padding; the distinction will be clear from context.) Computing $x \in \mathbb{C}^M$ then reduces to determining the first M entries of $x \in \ell(\mathbb{Z}_{4M-3})$ from $\mathrm{CirAut}(x + Rx)$. If x is completely real-valued, then this is indeed possible. For instance, consider the circular autocorrelation (11). If the entries of x are all real, then this becomes

$$\begin{aligned}\mathrm{CirAut}(x + Rx) = \big(&4a^2 + b^2 + c^2, 4ab + 2bc, b^2 + 4ac, 2bc, c^2,\\ &c^2, 2bc, b^2 + 4ac, 4ab + 2bc\big).\end{aligned}$$

Since $(\mathrm{CirAut}(x + Rx))(4) = c^2$, we simply take a square root to obtain c up to a sign. Assuming c is nonzero, we then divide $(\mathrm{CirAut}(x+Rx))(3)$ by 2c to determine b up to the same sign. Then subtracting b^2 from $(\mathrm{CirAut}(x + Rx))(2)$ and dividing by $4c$ gives a up to the same sign.

From this example, we see that the process of recovering the entries of x from $\mathrm{CirAut}(x + Rx)$ is iterative, working backward through its first $2M - 2$ entries. But what happens if c is zero? Fortunately, this process doesn't break: In this case, we have

$$\mathrm{CirAut}(x + Rx) = (4a^2 + b^2, 4ab, b^2, 0, 0, 0, 0, b^2, 4ab).$$

Thus, we need only start with $(\mathrm{CirAut}(x+Rx))(2)$ to determine the remaining entries of x up to a sign. This observation brings to light the important role of the last nonzero entry of x in the iteration. The relationship between this coordinate and the entries of $\mathrm{CirAut}(x+Rx)$ will become more rigorous later.

The above example illustrated how a real signal x is determined by $\mathrm{CirAut}(x+Rx)$. A complex-valued signal, on the other hand, is not completely determined from $\mathrm{CirAut}(x+Rx)$. Luckily, this can be fixed by introducing a second vector in $\ell(\mathbb{Z}_{4M-3})$ obtained from x, and we will demonstrate this later, but for now we focus on $x+Rx$. To this end, we first take a closer look at the entries of $\mathrm{CirAut}(x+Rx)$. Since this circular autocorrelation has even symmetry by construction, we need only consider all entries of $\mathrm{CirAut}(x+Rx)$ up to index $2M-2$. This leads to the following lemma:

Lemma 2.15. *Let x denote an M-dimensional complex signal embedded in $\ell(\mathbb{Z}_{4M-3})$ such that $x(p)=0$ for all $p=M,\ldots,4M-4$. Then*

$$(\mathrm{CirAut}(x+Rx))(p) = 2\,\mathrm{Re}\langle x, T^p x\rangle + \langle x, RT^{-p}x\rangle$$

for all $p=1,\ldots,2M-2$.

Proof. First note that by the definition of the circular autocorrelation in (9) we have

$$\begin{aligned}(\mathrm{CirAut}(x+Rx))(p) &= \langle x+Rx, T^p(x+Rx)\rangle \\ &= 2\,\mathrm{Re}\langle x, T^p x\rangle + \langle x, RT^{-p}x\rangle + \langle x, RT^p x\rangle.\end{aligned}$$

Thus, to complete the proof it suffices to show that $\langle x, RT^p x\rangle = 0$ for all $p = 1,\ldots,2M-2$. Since x is only nonzero in its first M entries, we have

$$\langle x, RT^p x\rangle = \sum_{p'=0}^{M-1} x(p')\overline{(RT^p x)(p')} = \sum_{p'=0}^{M-1} x(p')\overline{(T^p x)(-p')} = \sum_{p'=0}^{M-1} x(p')\overline{x(-p'-p)},$$

where the summand is zero whenever $-p'-p \notin [0, M-1]$ modulo $4M-3$. This is equivalent to having $-p$ not lie in the Minkowski sum $p' + [0, M-1]$, and since $p' \in [0, M-1]$ we see that $\langle x, RT^p x\rangle = 0$ for all $p = 1, \ldots, 2M-2$. □

As a consequence of Lemma 2.15, the following theorem expresses the entries of $\operatorname{CirAut}(x + Rx)$ in terms of the entries of x:

Theorem 2.16. *Let x denote an M-dimensional complex signal embedded in $\ell(\mathbb{Z}_{4M-3})$ such that $x(p) = 0$ for all $p = M, \ldots, 4M-4$. Then we have*

$$(\operatorname{CirAut}(x+Rx))(p) = \begin{cases} 2\operatorname{Re}\left(\displaystyle\sum_{p'=\frac{p+1}{2}}^{M-1} x(p')(\overline{x(p'-p)} + \overline{x(p-p')})\right) & \textit{if } p \textit{ is odd} \\ 2\operatorname{Re}\left(\displaystyle\sum_{p'=\frac{p}{2}+1}^{M-1} x(p')(\overline{x(p'-p)} + \overline{x(p-p')})\right) + \left|x(\tfrac{p}{2})\right|^2 & \textit{if } p \textit{ is even} \end{cases} \tag{12}$$

for all $p = 1, \ldots, 2M-2$.

Proof. We first use Lemma 2.15 to get

$$\begin{aligned}(\operatorname{CirAut}(x+Rx))(p) &= 2\operatorname{Re}\langle x, T^p x\rangle + \langle x, RT^{-p}x\rangle \\ &= 2\operatorname{Re}\left(\sum_{p'=0}^{M-1} x(p')\overline{x(p'-p)}\right) + \sum_{p'=0}^{M-1} x(p')\overline{x(p-p')} \\ &= 2\operatorname{Re}\left(\sum_{p'=p}^{M-1} x(p')\overline{x(p'-p)}\right) + \sum_{p'=\max\{p-(M-1),0\}}^{\min\{p,M-1\}} x(p')\overline{x(p-p')}, \end{aligned} \tag{13}$$

where the last equality takes into account that the first summand is nonzero only when $p'-p \in [0, M-1]$ and the second summand is nonzero only when $p-p' \in [0, M-1]$, i.e., when $p' \in [p, p+(M-1)]$ and $p' \in [p-(M-1), p]$, respectively. To continue, we divide our analysis into cases.

For $p = 1, \ldots, M-1$, (13) gives

$$(\mathrm{CirAut}(x + Rx))(p) = 2\,\mathrm{Re}\left(\sum_{p'=p}^{M-1} x(p')\overline{x(p'-p)}\right) + \sum_{p'=0}^{p} x(p')\overline{x(p-p')}. \tag{14}$$

If p is odd we can then write

$$\begin{aligned}\sum_{p'=0}^{p} x(p')\overline{x(p-p')} &= \sum_{p'=0}^{\frac{p-1}{2}} x(p')\overline{x(p-p')} + \sum_{p'=\frac{p+1}{2}}^{p} x(p')\overline{x(p-p')} \\ &= \sum_{p''=\frac{p+1}{2}}^{p} x(p-p'')\overline{x(p'')} + \sum_{p'=\frac{p+1}{2}}^{p} x(p')\overline{x(p-p')} \\ &= 2\,\mathrm{Re}\left(\sum_{p'=\frac{p+1}{2}}^{p} x(p')\overline{x(p-p')}\right), \end{aligned} \tag{15}$$

while if p is even we similarly write

$$\sum_{p'=0}^{p} x(p')\overline{x(p-p')} = 2\,\mathrm{Re}\left(\sum_{p'=\frac{p}{2}+1}^{p} x(p')\overline{x(p-p')}\right) + \left|x\left(\tfrac{p}{2}\right)\right|^2. \tag{16}$$

Substituting (15) and (16) into (14) then gives (12).

For the remaining case, $p = M, \ldots, 2M-2$ and (13) gives

$$(\mathrm{CirAut}(x + Rx))(p) = \sum_{p'=p-(M-1)}^{M-1} x(p')\overline{x(p-p')}. \tag{17}$$

Similar to the previous case, taking p to be odd yields

$$\sum_{p'=p-(M-1)}^{M-1} x(p')\overline{x(p-p')} = 2\,\mathrm{Re}\left(\sum_{p'=\frac{p+1}{2}}^{M-1} x(p')\overline{x(p-p')}\right), \tag{18}$$

while taking p to be even yields

$$\sum_{p'=p-(M-1)}^{M-1} x(p')\overline{x(p-p')} = 2\operatorname{Re}\left(\sum_{p'=\frac{p}{2}+1}^{M-1} x(p')\overline{x(p-p')}\right) + \left|x\left(\tfrac{p}{2}\right)\right|^2, \qquad (19)$$

and substituting (18) and (19) into (17) also gives (12). $\square$

Notice (12) shows that each member of $\{(\mathrm{CirAut}(x+Rx))(p)\}_{p=1}^{2M-2}$ can be written as a combination of the first M entries of x, but only those at or beyond the $\lceil\frac{p}{2}\rceil$th index. As such, the index of the last nonzero entry of x is closely related to that of the last nonzero entry of $\{(\mathrm{CirAut}(x+Rx))(p)\}_{p=1}^{2M-2}$. This corresponds to the observation earlier in the case of $x \in \mathbb{R}^3$ where the third coordinate was assumed to be zero. We identify the relationship between the locations of these nonzero entries in the following lemma:

Lemma 2.17. *Let x denote an M-dimensional complex signal embedded in $\ell(\mathbb{Z}_{4M-3})$ such that $x(p) = 0$ for all $p = M, \ldots, 4M-4$. Then the last nonzero entry of $\{(\mathrm{CirAut}(x+Rx))(p)\}_{p=0}^{2M-2}$ has index $p = 2q$, where q is the index of the last nonzero entry of x.*

Proof. If $q \geq 1$, then (12) gives that $(\mathrm{CirAut}(x+Rx))(2q) = |x(q)|^2 \neq 0$. Note that since $x(p') = 0$ for every $p' > q$, (12) also gives that $(\mathrm{CirAut}(x+Rx))(p) = 0$ for every $p > 2q$. For the remaining case where $q = 0$, (12) immediately gives that $(\mathrm{CirAut}(x+Rx))(p) = 0$ for every $p \geq 1$. To show that $(\mathrm{CirAut}(x+Rx))(0) \neq 0$ in this case, we apply the definition of circular autocorrelation in (9):

$$(\mathrm{CirAut}(x+Rx))(0) = \langle x+Rx, x+Rx\rangle = \|x+Rx\|^2 = |2x(0)|^2 \neq 0,$$

where the last equality uses the fact that x is only supported at 0 (since $q = 0$). $\square$

As previously mentioned, we are unable to recover the entries of a complex signal x solely from $\mathrm{CirAut}(x+Rx)$. One way to address this is to rotate the entries of x in the complex plane and also take the circular autocorrelation of this modified signal. If we rotate by an angle which is not an integer multiple of π, this will produce new entries which are linearly independent from the corresponding entries of x when viewed as vectors in the complex plane. As we will see, the problem of recovering the entries of x then reduces to solving a linear system.

Take any $(4M-3)\times(4M-3)$ diagonal modulation operator E whose diagonal entries $\{\omega_k\}_{k=0}^{4M-4}$ are of unit-modulus satisfying $\omega_j\overline{\omega_k}\notin\mathbb{R}$ for all $j\neq k$ and consider the new vector $Ex\in\ell(\mathbb{Z}_{4M-3})$. Then Theorem 2.16 gives

$$\begin{aligned}&(\mathrm{CirAut}(Ex+REx))(p)\\&=\begin{cases}2\operatorname{Re}\Bigg(\displaystyle\sum_{p'=\frac{p+1}{2}}^{M-1}\omega_{p'}x(p')(\overline{\omega_{p'-p}}\,\overline{x(p'-p)}+\overline{\omega_{p-p'}}\,\overline{x(p-p')})\Bigg) & \text{if } p \text{ is odd}\\ 2\operatorname{Re}\Bigg(\displaystyle\sum_{p'=\frac{p}{2}+1}^{M-1}\omega_{p'}x(p')(\overline{\omega_{p'-p}}\,\overline{x(p'-p)}+\overline{\omega_{p-p'}}\,\overline{x(p-p')})\Bigg)+\left|x(\tfrac{p}{2})\right|^2 & \text{if } p \text{ is even}\end{cases}\end{aligned}\tag{20}$$

for all $p=1,\dots,2M-2$. We will see that (12) and (20) together allow us to solve for the entries of x (up to a global phase factor) by working iteratively backward through the entries of $\mathrm{CirAut}(x+Rx)$ and $\mathrm{CirAut}(Ex+REx)$. As alluded to earlier, each entry index forms a linear system which can be solved using the following lemma:

Lemma 2.18. *Let $a,b\in\mathbb{C}\setminus\{0\}$ and $\omega\in\mathbb{C}\setminus\mathbb{R}$ with $|\omega|=1$. Then*

$$b=\frac{\mathrm{i}}{\overline{a}\,\mathrm{Im}(\omega)}\big(\operatorname{Re}(\omega a\overline{b})-\omega\operatorname{Re}(a\overline{b})\big).\tag{21}$$

Proof. By direct manipulation, we have

$$\begin{aligned}\operatorname{Re}(\omega a\bar{b}) - \omega \operatorname{Re}(a\bar{b}) &= \operatorname{Re}(\omega)\operatorname{Re}(a\bar{b}) - \operatorname{Im}(\omega)\operatorname{Im}(a\bar{b}) - \omega\operatorname{Re}(a\bar{b}) \\ &= -\mathrm{i}\operatorname{Im}(\omega)\big(\operatorname{Re}(a\bar{b}) - \mathrm{i}\operatorname{Im}(a\bar{b})\big) \\ &= -\mathrm{i}\operatorname{Im}(\omega)\big(\operatorname{Re}(\overline{a}b) + \mathrm{i}\operatorname{Im}(\overline{a}b)\big) \\ &= -\mathrm{i}\overline{a}b\operatorname{Im}(\omega).\end{aligned}$$

Rearranging then yields the desired result. □

We now use this lemma to describe how to recover x up to global phase. By Lemma 2.17, the last nonzero entry of $\{(\operatorname{CirAut}(x+Rx))(p)\}_{p=0}^{2M-2}$ has index $p = 2q$, where q indexes the last nonzero entry of x. As such, we know that $x(k) = 0$ for every $k > q$, and $x(q)$ can be estimated up to a phase factor ($\hat{x}(q) = e^{\mathrm{i}\psi}x(q)$) by taking the square root of $(\operatorname{CirAut}(x + Rx))(2q) = |x(q)|^2$ (we will verify this soon, but this corresponds to the examples we have seen so far). Next, if we know $\operatorname{Re}(x(q)\overline{x(k)})$ and $\operatorname{Re}(\omega_q\overline{\omega_k}x(q)\overline{x(k)})$ for some $k < q$, then we can use these to estimate $x(k)$:

$$\hat{x}(k) := \frac{\mathrm{i}}{\overline{\hat{x}(q)}\operatorname{Im}(\omega_q\overline{\omega_k})}\Big(\operatorname{Re}(\omega_q\overline{\omega_k}x(q)\overline{x(k)}) - \omega_q\overline{\omega_k}\operatorname{Re}(x(q)\overline{x(k)})\Big) = e^{\mathrm{i}\psi}x(k), \quad (22)$$

where the last equality follows from substituting $a = x(q)$, $b = x(k)$ and $\omega = \omega_q\overline{\omega_k}$ into (21). Overall, once we know $x(q)$ up to phase, we can then find $x(k)$ relative to this same phase for each $k = 0, \ldots, q-1$, provided we know $\operatorname{Re}(x(q)\overline{x(k)})$ and $\operatorname{Re}(\omega_q\overline{\omega_k}x(q)\overline{x(k)})$ for these k's. Thankfully, these values can be determined from the entries of $\operatorname{CirAut}(x + Rx)$ and $\operatorname{CirAut}(Ex + REx)$:

Theorem 2.19. *Let x denote an M-dimensional complex signal embedded in $\ell(\mathbb{Z}_{4M-3})$ such that $x(p) = 0$ for all $p = M, \ldots, 4M-4$ and E be a $(4M-3)\times(4M-3)$ diagonal modulation operator with diagonal entries $\{\omega_k\}_{k=0}^{4M-4}$ satisfying $|\omega_k| = 1$ for all $k = 0, \ldots, 4M-4$ and $\omega_j\overline{\omega_k} \notin \mathbb{R}$ for all $j \neq k$. Then x can be recovered up to a global phase factor from* $\operatorname{CirAut}(x + Rx)$ *and* $\operatorname{CirAut}(Ex + REx)$.

Proof. Letting q denote the index of the last nonzero entry of x, it suffices to estimate $\{x(k)\}_{k=0}^{q}$ up to a global phase factor. To this end, recall from Lemma 2.17 that the last nonzero entry of $\{(\mathrm{CirAut}(x+Rx))(p)\}_{p=0}^{2M-2}$ has index $p=2q$. If $q=0$, then we have already seen that $(\mathrm{CirAut}(x+Rx))(0)=4|x(0)|^2$. Since there exists $\psi\in[0,2\pi)$ such that $x(0)=e^{-\mathrm{i}\psi}|x(0)|$, we may take

$$\hat{x}(0):=\tfrac{1}{2}\sqrt{(\mathrm{CirAut}(x+Rx))(0)}=|x(0)|=e^{\mathrm{i}\psi}x(0).$$

Otherwise $q\in[1,M-1]$, and (12) gives

$$\begin{aligned}&(\mathrm{CirAut}(x+Rx))(2q)\\&\qquad=|x(q)|^2+2\,\mathrm{Re}\Bigg(\sum_{p'=q+1}^{M-1}x(p')(\overline{x(p'-2q)}+\overline{x(2q-p')})\Bigg)=|x(q)|^2.\end{aligned}$$

Thus, taking $\hat{x}(q):=\sqrt{(\mathrm{CirAut}(x+Rx))(2q)}=|x(q)|$ gives us $\hat{x}(q)=e^{\mathrm{i}\psi}x(q)$ for some $\psi\in[0,2\pi)$.

In the case where $q=1$, all that remains to determine is $\hat{x}(0)$, a calculation which we save for the end of the proof. For now, suppose $q\geq 2$. Since we already know $\hat{x}(q)=e^{\mathrm{i}\psi}x(q)$, we would like to determine $\hat{x}(k)$ for $k=1,\dots,q-1$. To this end, take $r\in[0,q-2]$ and suppose we have $\hat{x}(k)=e^{\mathrm{i}\psi}x(k)$ for all $k=q-r,\dots,q$. If we can obtain $\hat{x}(q-(r+1))$ up to the same phase from this information, then working iteratively from $r=0$ to $r=q-2$ will give us $\hat{x}(k)$ up to global phase for all but the zeroth entry (which we address later). Note when r is even, (12) gives

$$\begin{aligned}&(\mathrm{CirAut}(x+Rx))(2q-(r+1))\\&\qquad=2\,\mathrm{Re}\Bigg(\sum_{p'=q-\frac{r}{2}}^{q}x(p')(\overline{x(p'-(2q-(r+1)))}+\overline{x((2q-(r+1))-p')})\Bigg)\\&\qquad=2\,\mathrm{Re}\left(x(q)\overline{x(q-(r+1))}\right)+2\sum_{p'=q-\frac{r}{2}}^{q-1}\mathrm{Re}\left(x(p')\overline{x((2q-(r+1))-p')}\right),\end{aligned}$$

where the last equality follows from the observation that

$$p' - (2q - (r+1)) \leq -q + (r+1) \leq -1$$

over the range of the sum, meaning $x(p' - (2q - (r+1))) = 0$ throughout the sum. Similarly when r is odd, (12) gives

$$\begin{aligned}(\mathrm{CirAut}(x + Rx))(2q - (r+1)) = 2\operatorname{Re}\left(x(q)\overline{x(q-(r+1))}\right) + \left|x\left(q - \tfrac{r+1}{2}\right)\right|^2 \\ + 2\sum_{p'=q-\frac{r-1}{2}}^{q-1} \operatorname{Re}\left(x(p')\overline{x((2q-(r+1))-p')}\right).\end{aligned}$$

In either case, we can isolate $\operatorname{Re}(x(q)\overline{x(q-(r+1))})$ to get an expression in terms of $(\mathrm{CirAut}(x + Rx))(2q - (r+1))$ and other terms of the form $\operatorname{Re}(x(k)\overline{x(k')})$ or $|x(k)|^2$ for $k, k' \in [q-r, q-1]$. By the induction hypothesis, we have $\hat{x}(k) = e^{\mathrm{i}\psi}x(k)$ for $k = q-r, \ldots, q-1$, and so we can use these estimates to determine these other terms:

$$\begin{aligned}\operatorname{Re}(\hat{x}(k)\overline{\hat{x}(k')}) = \operatorname{Re}(e^{\mathrm{i}\psi}x(k)\overline{e^{\mathrm{i}\psi}x(k')}) = \operatorname{Re}(x(k)\overline{x(k')}), \\ |\hat{x}(k)|^2 = |e^{\mathrm{i}\psi}x(k)|^2 = |x(k)|^2.\end{aligned}$$

As such, we can use $(\mathrm{CirAut}(x + Rx))(2q - (r+1))$ along with the higher-indexed estimates $\hat{x}(k)$ to determine the term $\operatorname{Re}(x(q)\overline{x(q-(r+1))})$. Similarly, we can use $(\mathrm{CirAut}(Ex + REx))(2q - (r+1))$ along with the higher-indexed estimates $\hat{x}(k)$ to determine $\operatorname{Re}(\omega_q\overline{\omega_{(q-(r+1))}}x(q)\overline{x(q-(r+1))})$. We then plug these into (22), along with the estimate $\hat{x}(q) = e^{\mathrm{i}\psi}x(q)$ (which is also available by the induction hypothesis), to get $\hat{x}(2q - (r+1)) = e^{\mathrm{i}\psi}x(2q - (r+1))$.

At this point, we have determined $\{x(k)\}_{k=1}^{q}$ up to a global phase factor whenever $q \geq 1$, and so it remains to find $\hat{x}(0)$. For this, note that when q is odd, (12) gives

$$(\operatorname{CirAut}(x+Rx))(q) = 4\operatorname{Re}(x(q)\overline{x(0)}) + 2\sum_{p'=\frac{q+1}{2}}^{q-1} \operatorname{Re}\left(x(p')\overline{x(q-p')}\right),$$

while for even q, we have

$$(\operatorname{CirAut}(x+Rx))(q) = 4\operatorname{Re}(x(q)\overline{x(0)}) + 2\sum_{p'=\frac{q}{2}+1}^{q-1} \operatorname{Re}\left(x(p')\overline{x(q-p')}\right) + \left|x\left(\tfrac{q}{2}\right)\right|^2.$$

As before, isolating $\operatorname{Re}(x(q)\overline{x(0)})$ in either case produces an expression in terms of $(\operatorname{CirAut}(x+Rx))(q)$ and other terms of the form $\operatorname{Re}(x(k)\overline{x(k')})$ or $|x(k)|^2$ for $k, k' \in [1, q-1]$. These other terms can be calculated using the estimates $\{\hat{x}(k)\}_{k=1}^{q-1}$, and so we can also calculate $\operatorname{Re}(x(q)\overline{x(0)})$ from $(\operatorname{CirAut}(x+Rx))(q)$. Similarly, we can calculate $\operatorname{Re}(\omega_q\overline{\omega_0}x(q)\overline{x(0)})$ from $\{\hat{x}(k)\}_{k=1}^{q-1}$ and $(\operatorname{CirAut}(Ex+REx))(q)$, and plugging these into (22) along with $\hat{x}(q)$ produces the estimate $\hat{x}(0) = e^{\mathrm{i}\psi}x(0)$. □

Theorem 2.19 establishes that it is possible to recover a signal $x \in \mathbb{C}^M$ up to a global phase from $\{(\operatorname{CirAut}(x+Rx))(q)\}_{q=0}^{2M-2}$ and $\{(\operatorname{CirAut}(Ex+REx))(q)\}_{q=0}^{2M-2}$. We now return to how these circular autocorrelations relate to intensity measurements. Recall from (10) that the DFT of the circular autocorrelation is the modulus squared of the DFT of the original signal: $(F^*\operatorname{CirAut}(u))(q) = |(F^*u)(q)|^2$. Also note that the DFT commutes with the reversal operator:

$$\begin{aligned}(F^*Ru)(q) &= \sum_{p\in\mathbb{Z}_P} u(-p)e^{-2\pi ipq/P} = \sum_{p'\in\mathbb{Z}_P} u(p')e^{-2\pi ip'(-q)/P} \\ &= (F^*u)(-q) = (RF^*u)(q).\end{aligned}$$

With this, we can express $\operatorname{CirAut}(x+Rx)$ in terms of intensity measurements with a particular ensemble:

$$\begin{aligned}(F^*\operatorname{CirAut}(x+Rx))(q) &= |(F^*(x+Rx))(q)|^2 = |(F^*x)(q)+(F^*Rx)(q)|^2\\ &= |(F^*x)(q)+(F^*x)(-q)|^2 = |\langle x, f_q+f_{-q}\rangle|^2.\end{aligned}$$

Defining the qth *discrete cosine function* $c_q\in\ell(\mathbb{Z}_{4M-3})$ by

$$c_q(p) := 2\cos\left(\tfrac{2\pi pq}{4M-3}\right) = e^{2\pi \mathrm{i}pq/(4M-3)} + e^{-2\pi \mathrm{i}pq/(4M-3)} = (f_q+f_{-q})(p),$$

this means that $(F^*\operatorname{CirAut}(x+Rx))(q) = |\langle x, c_q\rangle|^2$ for all $q\in\mathbb{Z}_{4M-3}$. Similarly, if we take the modulation matrix E to have diagonal entries $\omega_k = e^{2\pi\mathrm{i}k/(2M-1)}$ for all $k=0,\ldots,4M-4$, we find

$$(F^*\operatorname{CirAut}(Ex+REx))(q) = |\langle Ex, c_q\rangle|^2 = |\langle x, E^*c_q\rangle|^2.$$

Thus, coupling the DFT with Theorem 2.19 allows us to recover the signal x from $4M-2$ intensity measurements, namely with the ensemble $\{c_q\}_{q=0}^{2M-2}\cup\{E^*c_q\}_{q=0}^{2M-2}$. Note that since $x\in\ell(\mathbb{Z}_{4M-3})$ is actually a zero-padded version of $x\in\mathbb{C}^M$, we may view c_q and E^*c_q as members of $\mathbb{C}^M$ by discarding the entries indexed by $p=M,\ldots,4M-4$.

Considering this section promised phase retrieval from only $4M-4$ intensity measurements, we must somehow find a way to discard two of these $4M-2$ measurement vectors. To do this, first note that

$$\begin{aligned}(\mathrm{CirAut}(Ex+REx))(0) &= \|Ex+REx\|^2\\ &= \sum_{k\in\mathbb{Z}_{4M-3}} \left|e^{2\pi\mathrm{i}k/(2M-1)}x(k)+e^{2\pi\mathrm{i}(-k)/(2M-1)}x(-k)\right|^2\\ &= \sum_{k=-(2M-2)}^{-1} \left|e^{2\pi\mathrm{i}(-k)/(2M-1)}x(-k)\right|^2+|2x(0)|^2\\ &\qquad+\sum_{k=1}^{2M-2}\left|e^{2\pi\mathrm{i}k/(2M-1)}x(k)\right|^2\\ &= \|x+Rx\|^2=\mathrm{CirAut}(x+Rx)(0).\end{aligned}$$

Moreover, we have

$$\begin{aligned}&(\mathrm{CirAut}(Ex+REx))(2M-2)\\ &\quad= \sum_{k\in\mathbb{Z}_{4M-3}}(Ex+REx)(k)\overline{(Ex+REx)(k-(2M-2))}\\ &\quad= (Ex+REx)(M-1)\overline{(Ex+REx)(-(M-1))}\\ &\quad= (Ex+REx)(M-1)\overline{(Ex+REx)(M-1)},\end{aligned}$$

where the last equality is by even symmetry. Since x is only supported on $k=0,\dots,M-1$, we then have

$$\begin{aligned}(\mathrm{CirAut}(Ex+REx))(2M-2) &= |(Ex+REx)(M-1)|^2\\ &= \left|e^{2\pi\mathrm{i}(M-1)/(2M-1)}x(M-1)+e^{-2\pi\mathrm{i}(M-1)/(2M-1)}x(-(M-1))\right|^2\\ &= \left|e^{2\pi\mathrm{i}(M-1)/(2M-1)}x(M-1)\right|^2\\ &= |x(M-1)|^2=(\mathrm{CirAut}(x+Rx))(2M-2).\end{aligned}$$

Furthermore, the even symmetry of the circular autocorrelation also gives

$$\begin{aligned}(\mathrm{CirAut}(Ex+REx))(-(2M-2)) &= (\mathrm{CirAut}(Ex+REx))(2M-2)\\ &= (\mathrm{CirAut}(x+Rx))(2M-2) = (\mathrm{CirAut}(x+Rx))(-(2M-2)).\end{aligned}$$

These redundancies between $\mathrm{CirAut}(x+Rx)$ and $\mathrm{CirAut}(Ex+REx)$ indicate that we might be able to remove measurement vectors from our ensemble while maintaining our ability to perform phase retrieval. The following theorem confirms this suspicion:

Theorem 2.20. *Let $c_q \in \mathbb{C}^M$ be the truncated discrete cosine function defined by $c_q(p) := 2\cos(\frac{2\pi pq}{4M-3})$ for all $p = 0, \ldots, M-1$, and let E be the $M \times M$ diagonal modulation operator with diagonal entries $\omega_k = e^{2\pi \mathrm{i} k/(2M-1)}$ for all $k = 0, \ldots, M-1$. Then the intensity measurement mapping $\mathcal{A} \colon \mathbb{C}^M/\mathbb{T} \to \mathbb{R}^{4M-4}$ defined by $\mathcal{A}(x) := \{|\langle x, c_q\rangle|^2\}_{q=0}^{2M-2} \cup \{|\langle x, E^* c_q\rangle|^2\}_{q=1}^{2M-3}$ is injective.*

Proof. Since Theorem 2.19 allows us to reconstruct any $x \in \mathbb{C}^M$ up to a global phase factor from the entries of $\mathrm{CirAut}(x+Rx)$ and $\mathrm{CirAut}(Ex+REx)$, it suffices to show that the intensity measurements $\{|\langle x, c_q\rangle|^2\}_{q=0}^{2M-2} \cup \{|\langle x, E^* c_q\rangle|^2\}_{q=1}^{2M-3}$ allow us to recover the entries of these circular autocorrelations. To this end, recall from (10) that these quantities are related through the inverse DFT:

$$\begin{aligned}\mathrm{CirAut}(x+Rx) &= (F^*)^{-1}\{|\langle x, c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}},\\ \mathrm{CirAut}(Ex+REx) &= (F^*)^{-1}\{|\langle x, E^* c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}}.\end{aligned}$$

Since we have $\{|\langle x, c_q\rangle|^2\}_{q=0}^{2M-2}$, we can exploit even symmetry to determine the rest of $\{|\langle x, c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}}$, and then apply the inverse DFT to get $\mathrm{CirAut}(x+Rx)$. Moreover, by the previous discussion, we also obtain the 0, $2M-2$, and $-(2M-2)$ entries of $\mathrm{CirAut}(Ex+REx)$ from the corresponding entries of $\mathrm{CirAut}(x+Rx)$. Organize this information about $\mathrm{CirAut}(Ex+REx)$ into a vector $w \in \ell(\mathbb{Z}_{4M-3})$ whose 0, $2M-2$, and $-(2M-2)$ entries come from $\mathrm{CirAut}(Ex+REx)$ and whose remaining entries are populated by even symmetry from $\{|\langle x, E^* c_q\rangle|^2\}_{q=1}^{2M-3}$. We can express w as a matrix-vector product $w = A\{|\langle x, E^* c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}}$, where A is the identity matrix with the 0, $2M-2$, and $-(2M-2)$ rows replaced by the corresponding rows of the inverse DFT matrix. To complete the proof, it suffices to show that the matrix A is invertible, since this would imply $\mathrm{CirAut}(Ex + REx) = (F^*)^{-1}A^{-1}w$.

Using the cofactor expansion, note that $\det(A)$ reduces to a determinant of a 3×3 submatrix of $(F^*)^{-1}$. Specifically, letting $\theta := 2\pi(2M-2)^2/(4M-3)$ we have

$$\det(A) = \det\left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{\mathrm{i}\theta} & e^{-\mathrm{i}\theta} \\ 1 & e^{-\mathrm{i}\theta} & e^{\mathrm{i}\theta} \end{bmatrix}\right) = (e^{2\mathrm{i}\theta} - e^{-2\mathrm{i}\theta}) - (e^{\mathrm{i}\theta} - e^{-\mathrm{i}\theta}) + (e^{-\mathrm{i}\theta} - e^{\mathrm{i}\theta})$$

$$= (e^{\mathrm{i}\theta} + e^{-\mathrm{i}\theta} - 2)(e^{\mathrm{i}\theta} - e^{-\mathrm{i}\theta}) = 4\mathrm{i}(\cos(\theta) - 1)\sin(\theta),$$

and so A is invertible if and only if $\cos(\theta) - 1 \neq 0$ and $\sin(\theta) \neq 0$. This is equivalent to having π not divide θ, and indeed, the ratio

$$\frac{\theta}{\pi} = \frac{2(2M-2)^2}{4M-3} = 2M - \frac{5}{2} + \frac{1}{2(4M-3)}$$

is not an integer because $M \geq 2$. As such, A is invertible. □

The following summarizes the measurement design and phase retrieval procedure described in this section:

Measurement design

- Define the qth truncated discrete cosine function $c_q := \{2\cos(\frac{2\pi pq}{4M-3})\}_{p=0}^{M-1}$
- Define the $M \times M$ diagonal matrix E with entries $\omega_k := e^{2\pi\mathrm{i}k/(2M-1)}$ for all $k = 0, \ldots, M-1$
- Take $\Phi := \{c_q\}_{q=0}^{2M-2} \cup \{E^* c_q\}_{q=1}^{2M-3}$

Phase retrieval procedure

- Calculate $\{|\langle x, c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}}$ from $\{|\langle x, c_q\rangle|^2\}_{q=0}^{2M-2}$ by even extension
- Calculate $\mathrm{CirAut}(x + Rx) = (F^*)^{-1}\{|\langle x, c_q\rangle|^2\}_{q\in\mathbb{Z}_{4M-3}}$
- Define $w \in \ell(\mathbb{Z}_{4M-3})$ so that its 0, $2M - 2$, and $-(2M - 2)$ entries are the corresponding entries in $\mathrm{CirAut}(x+Rx)$ and its remaining entries are populated by even symmetry from $\{|\langle x, E^*c_q\rangle|^2\}_{q=1}^{2M-3}$
- Define A to be the identity matrix with the 0, $2M - 2$, and $-(2M - 2)$ rows replaced by the corresponding rows of the inverse DFT matrix $(F^*)^{-1}$
- Calculate $\mathrm{CirAut}(Ex + REx) = (F^*)^{-1}A^{-1}w$
- Recover x up to global phase from $\mathrm{CirAut}(x + Rx)$ and $\mathrm{CirAut}(Ex + REx)$ using the process described in the proof of Theorem 2.19

III. Almost injective intensity measurements and the computational limits of phase retrieval

Now that we have a better understanding of injectivity in phase retrieval, it is natural to ask how much we might lose if we reduce the size of the measurement ensemble $\Phi = \{\varphi_n\}_{n=1}^N \subseteq V$, where $V = \mathbb{R}^M$ or $\mathbb{C}^M$, below the known and conjectured lower bounds ($2M-1$ for the real case and $4M-4$ for the complex case, respectively). Indeed, reducing the number of measurements is often desirable in practice as each measurement typically incurs some sort of cost. For instance, in the case of synthetic aperture radar, the number of measurements is proportional to the number of aircraft employed in the multistatic system, each of which contributes costs in energy and maintenance. Perhaps surprisingly, we can often decrease the number of measurements without losing much: As we will see, almost every signal can be completely determined from half the measurements required for injectivity. In this chapter, we address this issue by formally introducing a theory of almost injective intensity measurements, in which we relax the injectivity requirement to a set of signals that is dense in $\mathbb{R}^M$ ($\mathbb{C}^M$). For simplicity, we dedicate the analysis of almost injectivity to the real case, and we conclude by examining algorithmic efficiency in this setting.

3.1 Almost injectivity

While $4M + o(M)$ measurements are necessary and generically sufficient for injectivity in the complex case, one can save a factor of 2 in the number of measurements by slightly weakening the desired notion of injectivity [8, 57]. To be explicit, we start with the following definition:

Definition 3.1. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq V$, where $V = \mathbb{R}^M$ or $\mathbb{C}^M$. The intensity measurement mapping $\mathcal{A}\colon V/S \to \mathbb{R}^N$, where $S = \{\pm 1\}$ (resp. $\mathbb{T}$), defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$ is said to be almost injective if $\mathcal{A}^{-1}(\mathcal{A}(x)) = \{\omega x : |\omega| = 1\}$ for almost every $x \in V/S$.*

For the complex case, it is known that $2M$ measurements are necessary for almost injectivity [57], and that $2M$ generic measurements suffice [8] (cf. [56]); this is the factor-of-2 savings mentioned above. For the real case, it is also known how many measurements are necessary and generically sufficient for almost injectivity: $M+1$ [8]. Like the complex case, this is also a factor-of-2 savings from the injectivity requirement: $2M-1$. This requirement for injectivity in the real case follows from the complement property characterization of injectivity from [8] (Theorem 2.2 of this paper). Similar to this result, we will characterize ensembles of measurement vectors which yield almost injective intensity measurements and, similar to its proof, the basic idea behind the analysis is to consider sums and differences of signals with identical intensity measurements. However, the characterization we develop is limited to the real case; a similar analysis for the complex case remains an open problem.

Lemma 3.2. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ and the intensity measurement mapping $\mathcal{A}\colon \mathbb{R}^M/\{\pm 1\} \rightarrow \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$. Then $\mathcal{A}$ is almost injective if and only if almost every $x \in \mathbb{R}^M$ is not in the Minkowski sum $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for all $S \subseteq \{1,\ldots,N\}$. More precisely, $\mathcal{A}^{-1}(\mathcal{A}(x)) = \{\pm x\}$ if and only if $x \notin \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for any $S \subseteq \{1,\ldots,N\}$.*

Proof. By the definition of the mapping $\mathcal{A}$, for $x, y \in \mathbb{R}^M$ we have $\mathcal{A}(x) = \mathcal{A}(y)$ if and only if $|\langle x, \varphi_n\rangle| = |\langle y, \varphi_n\rangle|$ for all $n \in \{1,\ldots,N\}$. This occurs precisely when there is a subset $S \subseteq \{1,\ldots,N\}$ such that $\langle x, \varphi_n\rangle = -\langle y, \varphi_n\rangle$ for every $n \in S$ and $\langle x, \varphi_n\rangle = \langle y, \varphi_n\rangle$ for every $n \in S^c$. Thus, $\mathcal{A}^{-1}(\mathcal{A}(x)) = \{\pm x\}$ if and only if for every $y \neq \pm x$ and for every $S \subseteq \{1,\ldots,N\}$, either there exists an $n \in S$ such that $\langle x+y, \varphi_n\rangle \neq 0$ or an $n \in S^c$ such that $\langle x-y, \varphi_n\rangle \neq 0$. We claim that this occurs if and only if x is not in the Minkowski sum $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for all $S \subseteq \{1,\ldots,N\}$, which would complete the proof. We verify the claim by seeking the contrapositive in each direction.

($\Rightarrow$) Suppose $x \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$. Then there exists $u \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $v \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ such that $x = u+v$. Taking $y := u-v$, we see that $x+y = 2u \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $x-y = 2v \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$, which means that there is no $n \in S$ such that $\langle x+y, \varphi_n \rangle \neq 0$ nor $n \in S^c$ such that $\langle x-y, \varphi_n \rangle \neq 0$. Furthermore, u and v are nonzero, and so $y \neq \pm x$.

($\Leftarrow$) Suppose $y \neq \pm x$ and for every $S \subseteq \{1, \ldots, N\}$ there is no $n \in S$ such that $\langle x+y, \varphi_n \rangle \neq 0$ nor $n \in S^c$ such that $\langle x-y, \varphi_n \rangle \neq 0$. Then $x+y \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\}$ and $x-y \in \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$. Since $x = \frac{1}{2}(x+y) + \frac{1}{2}(x-y)$, we have that $x \in \operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$. □

If the mapping $\mathcal{A}$ is injective, then the ensemble Φ in Lemma 3.2 must satisfy the complement property, and so the set $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ is of the form $V \setminus \{0\} + \emptyset = \emptyset$ regardless of the choice of $S \subseteq \{1, \ldots, N\}$ (here, $V \subseteq \mathbb{R}^M$ is a proper subspace). Hence, this Minkowski sum requirement slightly weakens what it means to be injective. We will continue to investigate this set with the aid of the following lemma:

Lemma 3.3. *Let U and V be subspaces of a common vector space. If $U \cap V = \{0\}$, then $U \setminus \{0\} + V \setminus \{0\} = (U+V) \setminus (U \cup V)$.*

Proof. Since $U \setminus \{0\} + V \setminus \{0\}$ is a subset of $U+V$, it suffices to show that $(U \setminus \{0\} + V \setminus \{0\}) \cap (U \cup V) = \emptyset$. To this end, suppose $x \in U \setminus \{0\} + V \setminus \{0\}$. Then $x = u+v$ for some nonzero vectors $u \in U$ and $v \in V$. Since $U \cap V = \{0\}$, it follows that $x \notin U$ and $x \notin V$, that is, $x \notin U \cup V$. Likewise, if $x \in U \cup V$, then the fact that $U \cap V = \{0\}$ implies $x = u+v$ for some $u \in U$ and $v \in V$ where either u or v is zero. Hence, $x \notin U \setminus \{0\} + V \setminus \{0\}$, completing the proof. □

Theorem 3.4. *Consider $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ and the intensity measurement mapping $\mathcal{A} \colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ defined by $(\mathcal{A}(x))(n) := |\langle x, \varphi_n \rangle|^2$. Suppose Φ spans $\mathbb{R}^M$ and each φ_n is nonzero. Then $\mathcal{A}$ is almost injective if and only if the Minkowski* *sum* $\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp$ *is a proper subspace of* $\mathbb{R}^M$ *for each nonempty proper subset* $S \subseteq \{1, \ldots, N\}$.

Note that this result is not terribly surprising considering Lemma 3.2, as the new condition involves a simpler Minkowski sum in exchange for additional (reasonable and testable) assumptions on Φ. The proof of this theorem amounts to measuring the difference between the two Minkowski sums:

Proof of Theorem 3.4. First note that the spanning assumption on Φ implies

$$\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp = \big(\operatorname{span}(\Phi_S) + \operatorname{span}(\Phi_{S^c})\big)^\perp = \operatorname{span}(\Phi)^\perp = \{0\},$$

and so applying Lemma 3.3 yields the following identity:

$$\begin{aligned} &\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\} \\ &\qquad = \big(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) \setminus \big(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\big). \end{aligned} \tag{23}$$

From Lemma 3.2 we know that $\mathcal{A}$ is almost injective if and only if almost every $x \in \mathbb{R}^M$ is not in the Minkowski sum $\operatorname{span}(\Phi_S)^\perp \setminus \{0\} + \operatorname{span}(\Phi_{S^c})^\perp \setminus \{0\}$ for any $S \subseteq \{1, \ldots, N\}$. In other words, the Lebesgue measure (which we denote by $\operatorname{Leb}[\cdot]$) of this Minkowski sum is zero for each $S \subseteq \{1, \ldots, N\}$. By (23), this equivalently means that the Lebesgue measure of $\big(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) \setminus \big(\operatorname{span}(\Phi_S)^\perp \cup \operatorname{span}(\Phi_{S^c})^\perp\big)$ is zero for each $S \subseteq \{1, \ldots, N\}$. Since Φ spans $\mathbb{R}^M$, this set is empty (and therefore has Lebesgue measure zero) when $S = \emptyset$ or $S = \{1, \ldots, N\}$. Also, since each φ_n is nonzero, we know that $\operatorname{span}(\Phi_S)^\perp$ and $\operatorname{span}(\Phi_{S^c})^\perp$ are proper subspaces of $\mathbb{R}^M$ whenever S is a nonempty proper subset of $\{1, \ldots, N\}$, and so in these cases both subspaces must have Lebesgue measure zero. As such, we have that for every nonempty proper subset $S \subseteq \{1,\ldots,N\}$,

$$\begin{aligned}\mathrm{Leb}&\left[\big(\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp\big)\setminus\big(\mathrm{span}(\Phi_S)^\perp\cup\mathrm{span}(\Phi_{S^c})^\perp\big)\right]\\ &\geq \mathrm{Leb}\left[\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp\right]-\mathrm{Leb}\left[\mathrm{span}(\Phi_S)^\perp\right]-\mathrm{Leb}\left[\mathrm{span}(\Phi_{S^c})^\perp\right]\\ &= \mathrm{Leb}\left[\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp\right]\\ &\geq \mathrm{Leb}\left[\big(\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp\big)\setminus\big(\mathrm{span}(\Phi_S)^\perp\cup\mathrm{span}(\Phi_{S^c})^\perp\big)\right].\end{aligned}$$

In summary, $\big(\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp\big)\setminus\big(\mathrm{span}(\Phi_S)^\perp\cup\mathrm{span}(\Phi_{S^c})^\perp\big)$ having Lebesgue measure zero for each $S\subseteq\{1,\ldots,N\}$ is equivalent to $\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp$ having Lebesgue measure zero for each nonempty proper subset $S\subseteq\{1,\ldots,N\}$, which in turn is equivalent to the Minkowski sum $\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp$ being a proper subspace of $\mathbb{R}^M$ for each nonempty proper subset $S\subseteq\{1,\ldots,N\}$, as desired. □

At this point, consider the following stronger restatement of Theorem 3.4: "Suppose each φ_n is nonzero. Then $\mathcal{A}$ is almost injective if and only if Φ spans $\mathbb{R}^M$ and the Minkowski sum $\mathrm{span}(\Phi_S)^\perp+\mathrm{span}(\Phi_{S^c})^\perp$ is a proper subspace of $\mathbb{R}^M$ for each nonempty proper subset $S\subseteq\{1,\ldots,N\}$." Note that we can move the spanning assumption into the condition because if Φ does not span, then we can decompose almost every $x\in\mathbb{R}^M$ as $x=u+v$ such that $u\in\mathrm{span}(\Phi)$ and $v\in\mathrm{span}(\Phi)^\perp$ with $v\neq 0$, and defining $y:=u-v$ then gives $\mathcal{A}(y)=\mathcal{A}(x)$ despite the fact that $y\neq\pm x$. As for the assumption that the φ_n's are nonzero, we note that having $\varphi_n=0$ amounts to having the nth entry of $\mathcal{A}(x)$ be zero for all x. As such, Φ yields almost injectivity precisely when the nonzero members of Φ together yield almost injectivity. With this identification, the stronger restatement of Theorem 3.4 above can be viewed as a complete characterization of almost injectivity. Next, we will replace the Minkowski sum condition with a rather elegant condition involving the ranks of Φ_S and Φ_{S^c} by applying the following lemma:

Lemma 3.5 (Inclusion-exclusion principle for subspaces)**.** *Let U and V be subspaces of a common vector space. Then* $\dim(U+V)=\dim U+\dim V-\dim(U\cap V)$.

Proof. Let A be a basis for $U \cap V$ and let B and C be bases for U and V, respectively, such that $A \subseteq B$ and $A \subseteq C$. It can be shown that $A \cup B \cup C$ forms a basis for $U + V$, which implies that

$$\dim(U+V) = |A|+|B \setminus A|+|C \setminus A| = |B|+|C|-|A| = \dim U+\dim V-\dim(U \cap V),$$

completing the proof. $\square$

Theorem 3.6. *Consider* $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ *and the intensity measurement mapping* $\mathcal{A}\colon \mathbb{R}^M/\{\pm 1\} \to \mathbb{R}^N$ *defined by* $(\mathcal{A}(x))(n) := |\langle x, \varphi_n\rangle|^2$*. Suppose each* φ_n *is nonzero. Then* $\mathcal{A}$ *is almost injective if and only if* Φ *spans* $\mathbb{R}^M$ *and* $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} > M$ *for each nonempty proper subset* $S \subseteq \{1, \dots, N\}$*.*

Proof. Considering the discussion after the proof of Theorem 3.4, it suffices to assume that Φ spans $\mathbb{R}^M$. Furthermore, considering Theorem 3.4, it suffices to characterize when $\dim\big(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) < M$. By Lemma 3.5, we have

$$\begin{aligned}
&\dim\big(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) \\
&\qquad = \dim\big(\operatorname{span}(\Phi_S)^\perp\big) + \dim\big(\operatorname{span}(\Phi_{S^c})^\perp\big) - \dim\big(\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp\big).
\end{aligned}$$

Since Φ is assumed to span $\mathbb{R}^M$, we also have that $\operatorname{span}(\Phi_S)^\perp \cap \operatorname{span}(\Phi_{S^c})^\perp = \{0\}$, and so

$$\begin{aligned}
\dim\big(&\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) \\
&= \Big(M - \dim\left(\operatorname{span}(\Phi_S)\right)\Big) + \Big(M - \dim\left(\operatorname{span}(\Phi_{S^c})\right)\Big) - 0 \\
&= 2M - \operatorname{rank}\Phi_S - \operatorname{rank}\Phi_{S^c}.
\end{aligned}$$

As such, $\dim\big(\operatorname{span}(\Phi_S)^\perp + \operatorname{span}(\Phi_{S^c})^\perp\big) < M$ precisely when $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} > M$. $\square$

At this point, we point out some interesting consequences of Theorem 3.6. First of all, Φ cannot be almost injective if $N < M+1$ since $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} \leq |S| + |S^c| = N$. Also, in the case where $N = M+1$, we note that Φ is almost injective precisely when Φ is full spark, that is, every size-M subcollection is a spanning set (note this implies that all of the φ_n's are nonzero). In fact, every full spark Φ with $N \geq M+1$ yields almost injective intensity measurements, which in turn implies that a generic Φ yields almost injectivity when $N \geq M+1$ [8]. This is in direct analogy with injectivity in the real case; here, injectivity requires $N \geq 2M-1$, injectivity with $N = 2M-1$ is equivalent to being full spark, and being full spark suffices for injectivity whenever $N \geq 2M-1$ [8]. Another thing to check is that the condition for injectivity implies the condition for almost injectivity: Since the mapping $\mathcal{A}$ is injective for real Φ if and only if Φ is CP, it follows that $\operatorname{rank}\Phi_S + \operatorname{rank}\Phi_{S^c} \geq M+1 > M$ for every nonempty proper subset $S \subseteq \{1,\ldots,N\}$.

Having established that full spark ensembles of size $N \geq M+1$ yield almost injective intensity measurements, we note that checking whether a matrix is full spark is NP-hard in general [70]. Granted, there are a few explicit constructions of full spark ensembles which can be used [4, 81], but it would be nice to have a condition which is not computationally difficult to test in general. We provide one such condition in the next theorem, but first, we briefly review the requisite frame theory.

A *frame* is an ensemble $\Phi = \{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$ together with *frame bounds* $0 < A \leq B < \infty$ with the property that for every $x \in \mathbb{R}^M$,

$$A\|x\|^2 \leq \sum_{n=1}^N |\langle x, \varphi_n\rangle|^2 \leq B\|x\|^2.$$

When $A = B$, the frame is said to be *tight*, and such frames come with a painless reconstruction formula:

$$x = \frac{1}{A}\sum_{n=1}^N \langle x, \varphi_n\rangle\varphi_n.$$

To be clear, the theory of frames originated in the context of infinite-dimensional Hilbert spaces [41,46], and frames have since been studied in finite-dimensional settings, primarily because this is the setting in which they are applied computationally. Of particular interest are so-called *unit norm tight frames (UNTFs)*, which are tight frames whose frame elements have unit norm: $\|\varphi_n\| = 1$ for every $n = 1, \ldots, N$. Such frames are useful in applications; for example, if one encodes a signal x using frame coefficients $\langle x, \varphi_n \rangle$ and transmits these coefficients across a channel, then UNTFs are optimally robust to noise [58] and one erasure [32]. Intuitively, this optimality comes from the fact that frame elements of a UNTF are particularly well-distributed in the unit sphere [14]. Another pleasant feature of UNTFs is that it is straightforward to test whether a given frame is a UNTF: Letting $\Phi = [\varphi_1 \cdots \varphi_N]$ denote an $M \times N$ matrix whose columns are the frame elements, then Φ is a UNTF precisely when each of the following occurs simultaneously:

(i) the rows have equal norm

(ii) the rows are orthogonal

(iii) the columns have unit-norm

(This is a direct consequence of the tight frame's reconstruction formula and the fact that a UNTF has unit-norm frame elements; furthermore, since the columns have unit-norm, it is not difficult to see that the rows will necessarily have norm $\sqrt{N/M}$.) In addition to being able to test that an ensemble is a UNTF, various UNTFs can be constructed using *spectral tetris* [30] (though such frames necessarily have $N \geq 2M$), and *every* UNTF can be constructed using the recent theory of *eigensteps* [20,53]. Now that UNTFs have been properly introduced, we relate them to almost injectivity for phase retrieval:

Theorem 3.7. *If M and N are relatively prime, then every unit norm tight frame $\Phi = \{\varphi_n\}_{n=1}^{N} \subseteq \mathbb{R}^M$ yields almost injective intensity measurements.*

Proof. Pick a nonempty proper subset $S \subseteq \{1, \ldots, N\}$. By Theorem 3.6, it suffices to show that $\operatorname{rank} \Phi_S + \operatorname{rank} \Phi_{S^c} > M$, or equivalently, $\operatorname{rank} \Phi_S\Phi_S^* + \operatorname{rank} \Phi_{S^c}\Phi_{S^c}^* > M$. Note that since Φ is a unit norm tight frame, we also have

$$\Phi_S\Phi_S^* + \Phi_{S^c}\Phi_{S^c}^* = \Phi\Phi^* = \tfrac{N}{M}I,$$

and so $\Phi_S\Phi_S^*$ and $\Phi_{S^c}\Phi_{S^c}^*$ are simultaneously diagonalizable, i.e., there exists a unitary matrix U and diagonal matrices D_1 and D_2 such that

$$UD_1U^* + UD_2U^* = \Phi_S\Phi_S^* + \Phi_{S^c}\Phi_{S^c}^* = \tfrac{N}{M}I.$$

Conjugating by U^*, this then implies that $D_1 + D_2 = \frac{N}{M}I$. Let $L_1 \subseteq \{1, \ldots, M\}$ denote the diagonal locations of the nonzero entries in D_1, and $L_2 \subseteq \{1, \ldots, M\}$ similarly for D_2. To complete the proof, we need to show that $|L_1| + |L_2| > M$ (since $|L_1| + |L_2| = \operatorname{rank} \Phi_S\Phi_S^* + \operatorname{rank} \Phi_{S^c}\Phi_{S^c}^*$). Note that $L_1 \cup L_2 \neq \{1, \ldots, M\}$ would imply that $D_1 + D_2$ has at least one zero in its diagonal, contradicting the fact that $D_1 + D_2$ is a nonzero multiple of the identity; as such, $L_1 \cup L_2 = \{1, \ldots, M\}$ and $|L_1| + |L_2| \geq M$. We claim that this inequality is strict due to the assumption that M and N are relatively prime. To see this, it suffices to show that $L_1 \cap L_2$ is nonempty. Suppose to the contrary that L_1 and L_2 are disjoint. Then since $D_1 + D_2 = \frac{N}{M}I$, every nonzero entry in D_1 must be N/M. Since S is a nonempty proper subset of $\{1, \ldots, N\}$, this means that there exists $K \in (0, M)$ such that D_1 has K entries which are N/M and $M - K$ which are 0. Thus,

$$|S| = \operatorname{Tr}[\Phi_S^*\Phi_S] = \operatorname{Tr}[\Phi_S\Phi_S^*] = \operatorname{Tr}[UD_1U^*] = \operatorname{Tr}[D_1] = K(N/M),$$

implying that $N/M = |S|/K$ with $K \neq M$ and $|S| \neq N$. Since this contradicts the assumption that N/M is in lowest form, we have the desired result. $\square$

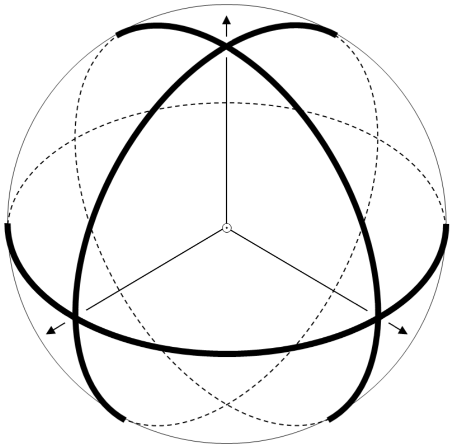


Figure 2: The simplex in $\mathbb{R}^3$. Pointing out of the page is the vector $\frac{1}{\sqrt{3}}(1,1,1)$, while the other vectors are the three permutations of $\frac{1}{\sqrt{3}}(1,-1,-1)$. Together, these four vectors form a unit norm tight frame, and since $M=3$ and $N=4$ are relatively prime, these yield almost injective intensity measurements in accordance with Theorem 3.7. For this ensemble, the points x such that $\mathcal{A}^{-1}(\mathcal{A}(x)) \neq \{\pm x\}$ are contained in the three coordinate planes. Above, we depict the intersection between these planes and the unit sphere. According to Theorem 3.9, performing phase retrieval with simplices such as this is NP-hard.

In general, whether a UNTF Φ yields almost injective intensity measurements is determined by whether it is *orthogonally partitionable*: Φ is orthogonally partitionable if there exists a partition $S \sqcup S^\mathrm{c} = \{1,\ldots,N\}$ such that $\operatorname{span}(\Phi_S)$ is orthogonal to $\operatorname{span}(\Phi_{S^\mathrm{c}})$. Specifically, a UNTF yields almost injective intensity measurements precisely when it is not orthogonally partitionable. Historically, this property of UNTFs has been pivotal to the understanding of singularities in the algebraic variety of UNTFs [47], and it has also played a key role in solutions to the Paulsen problem [16, 29]. However, it is not clear in general how to efficiently test for this property; this is why Theorem 3.7 is so powerful.

3.2 The computational complexity of phase retrieval

The previous section characterized the real ensembles which yield almost injective intensity measurements. The benefit of seeking almost injectivity instead of injectivity is that we can get away with much smaller ensembles. For example, Theorem 3.7 implies that a full spark ensemble in $\mathbb{R}^M$ of size $M+1$ suffices for almost injectivity, while $2M-1$ measurements are required for injectivity (Theorem 2.2).

In this section, we demonstrate that this savings in the number of measurements can come at a substantial price in computational requirements for phase retrieval. In particular, we consider the following problem:

Problem 3.8. *Let $\mathcal{F} = \{\Phi_M\}_{M=2}^{\infty}$ be a family of ensembles $\Phi_M = \{\varphi_{M;n}\}_{n=1}^{N(M)} \subseteq \mathbb{R}^M$, where $N(M) = \mathsf{poly}(M)$. Then* ConsistentIntensities$[\mathcal{F}]$ *is the following problem: Given $M \geq 2$ and a rational sequence $\{b_n\}_{n=1}^{N(M)}$, does there exist $x \in \mathbb{R}^M$ such that $|\langle x, \varphi_{M;n}\rangle| = b_n$ for every $n = 1, \dots, N(M)$?*

In this section, we will evaluate the computational complexity of the problem ConsistentIntensities$[\mathcal{F}]$ for a large class of families of small ensembles $\mathcal{F}$, but first, we briefly review the main concepts involved. Complexity theory is chiefly concerned with *complexity classes*, which are sets of problems that share certain computational requirements, such as time or space. For example, the complexity class P is the set of problems which can be solved in an amount of time that is bounded by some polynomial of the bit-length of the input. As another example, NP contains all problems for which an affirmative answer comes with a certificate that can be verified in polynomial time; note that $\mathsf{P} \subseteq \mathsf{NP}$ since for every problem $A \in \mathsf{P}$, one may ignore the certificate and find the affirmative answer in polynomial time. One key tool that is used to evaluate the complexity of a problem is called *polynomial-time reduction.* This is a polynomial-time algorithm that solves a problem A by exploiting an oracle which solves another problem B, indicating that solving A is no harder than solving B (up to polynomial factors in time); if such a reduction exists, we write $A \leq B$. For example, any efficient phase retrieval procedure for $\mathcal{F}$ can be used as a subroutine to solve ConsistentIntensities$[\mathcal{F}]$, indicating that phase retrieval for $\mathcal{F}$ is at least as hard as ConsistentIntensities$[\mathcal{F}]$. A problem B is called NP-*hard* if $B \geq A$ for every problem $A \in \mathsf{NP}$. Note that since $\leq$ is transitive, it suffices to show that $B \geq C$ for some NP-hard problem C. Finally, a problem B is called NP-*complete* if $B \in \mathsf{NP}$ is NP-hard; intuitively, NP-complete problems are the hardest of problems in NP. It is an open problem whether $\mathsf{P} = \mathsf{NP}$, but inequality is widely believed [39]; note that under this assumption, NP-hard problems have no computationally efficient solution. This provides a proper context for the main result of this section:

Theorem 3.9. *Let* $\mathcal{F} = \{\Phi_M\}_{M=2}^{\infty}$ *denote a family of full spark ensembles* $\Phi_M = \{\varphi_{M;n}\}_{n=1}^{M+1} \subseteq \mathbb{R}^M$ *with rational entries that can be computed in polynomial time. Then* ConsistentIntensities$[\mathcal{F}]$ *is* NP*-complete.*

Note that since the ensembles Φ_M are full spark, the existence of a solution to the phase retrieval problem $|\langle x, \varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M+1$ implies uniqueness by Theorem 3.6. Before proving this theorem, we first relate it to a previous hardness result from [82]. Specifically, this result can be restated using the terminology in this paper as follows: There exists a family $\mathcal{F} = \{\Phi_M\}_{M=2}^{\infty}$ of ensembles $\Phi_M = \{\varphi_{M;n}\}_{n=1}^{2M} \subseteq \mathbb{C}^M$, each of which yielding almost injective intensity measurements, such that ConsistentIntensities$[\mathcal{F}]$ is NP-complete. Interestingly, these are the smallest possible almost injective ensembles in the complex case, and we suspect that the result can be strengthened to the obvious analogy of Theorem 3.9:

Conjecture 3.10. *Let* $\mathcal{F} = \{\Phi_M\}_{M=2}^{\infty}$ *be a family of ensembles* $\Phi_M = \{\varphi_{M;n}\}_{n=1}^{2M} \subseteq \mathbb{C}^M$ *which yield almost injective intensity measurements and have complex rational entries that can be computed in polynomial time. Then* ConsistentIntensities$[\mathcal{F}]$ *is* NP*-complete.*

To prove Theorem 3.9, we devise a polynomial-time reduction from the following problem which is well-known to be NP-complete [68]:

Problem 3.11 (SubsetSum)**.** *Given a finite collection of integers* A *and an integer* z*, does there exist a subset* $S \subseteq A$ *such that* $\sum_{a \in S} a = z$*?*

Proof of Theorem 3.9. We first show that ConsistentIntensities$[\mathcal{F}]$ is in NP. Note that if there exists an $x \in \mathbb{R}^M$ such that $|\langle x, \varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M+1$, then x will have all rational entries. Indeed, $v := \Phi_M^* x$ has all rational entries, being a signed version of $\{b_n\}_{n=1}^{M+1}$, and so $x = (\Phi_M \Phi_M^*)^{-1} \Phi_M v$ is also rational. Thus, we can view x as a certificate of finite bit-length, and for each $n = 1, \ldots, M+1$, we know that $|\langle x, \varphi_{M;n}\rangle| = b_n$ can be verified in time which is polynomial in this bit-length, as desired.

Now we show that $\textsc{ConsistentIntensities}[\mathcal{F}]$ is NP-hard by reduction from $\textsc{SubsetSum}$. To this end, take a finite collection of integers A and an integer z. Set $M := |A|$ and label the members of A as $\{a_m\}_{m=1}^M$. Let Ψ denote the $M \times M$ matrix whose columns are the first M members of Φ_M. Since Φ_M is full spark, Ψ is invertible and $\Psi^{-1}\Phi_M$ has the form $[I\ w]$, where w has all nonzero entries; indeed, if the mth entry of w were zero, then $\Phi_M \setminus \{\varphi_{M;m}\}$ would not span, violating the full spark property of Φ_M. Now define

$$b_n := \begin{cases} \left|\dfrac{a_n}{w_n}\right| & \text{if } n = 1, \ldots, M \\ \left|2z - \displaystyle\sum_{m=1}^{M} a_m\right| & \text{if } n = M+1. \end{cases} \tag{24}$$

We claim that an oracle for $\textsc{ConsistentIntensities}[\mathcal{F}]$ would return “yes” from the inputs M and $\{b_n\}_{n=1}^{M+1}$ defined above if and only if there exists a subset $S \subseteq A$ such that $\sum_{a \in S} a = z$, which would complete the reduction.

To prove this claim, we start with $(\Rightarrow)$: Suppose there exists $x \in \mathbb{R}^M$ such that $|\langle x, \varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M+1$. Then $y := \Psi^* x$ satisfies $|\langle y, \Psi^{-1}\varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M+1$. Since $\Psi^{-1}\Phi_M = [I\ w]$, then by (24), the entries of y satisfy

$$|y_m| = \left|\frac{a_m}{w_m}\right| \quad \forall m = 1, \ldots, M \qquad \text{and} \qquad \left|\sum_{m=1}^{M} y_m w_m\right| = \left|2z - \sum_{m=1}^{M} a_m\right|.$$

By the first equation above, there exists a sequence $\{\varepsilon_m\}_{m=1}^M$ of ±1's such that $y_m = \varepsilon_m a_m / w_m$ for every $m = 1, \ldots, M$, and so the second equation above gives

$$\begin{aligned}\bigg|2z - \sum_{m=1}^M a_m\bigg| = \bigg|\sum_{m=1}^M y_m w_m\bigg| &= \bigg|\sum_{m=1}^M \varepsilon_m a_m\bigg| \\ &= \bigg|\sum_{\substack{m=1\\ \varepsilon_m=1}}^M a_m - \sum_{\substack{m=1\\ \varepsilon_m=-1}}^M a_m\bigg| = \bigg|2\sum_{\substack{m=1\\ \varepsilon_m=1}}^M a_m - \sum_{m=1}^M a_m\bigg|.\end{aligned}$$

Removing the absolute values, this means the left-hand side above is equal to the right-hand side, up to a sign factor. At this point, isolating z reveals that $z = \sum_{m\in S} a_m$, where S is either $\{m : \varepsilon_m = 1\}$ or $\{m : \varepsilon_m = -1\}$, depending on the sign factor.

For ($\Leftarrow$), suppose there is a subset $S \subseteq \{1, \ldots, M\}$ such that $z = \sum_{m\in S} a_m$. Define $\varepsilon_m := 1$ when $m \in S$ and $\varepsilon_m := -1$ when $m \notin S$. Then

$$\bigg|\sum_{m=1}^M \varepsilon_m a_m\bigg| = \bigg|\sum_{\substack{m=1\\ \varepsilon_m=1}}^M a_m - \sum_{\substack{m=1\\ \varepsilon_m=-1}}^M a_m\bigg| = \bigg|2\sum_{\substack{m=1\\ \varepsilon_m=1}}^M a_m - \sum_{m=1}^M a_m\bigg| = \bigg|2z - \sum_{m=1}^M a_m\bigg|.$$

By the analysis from the ($\Rightarrow$) direction, taking $y_m := \varepsilon_m a_m / w_m$ for each $m = 1, \ldots, M$ then ensures that $|\langle y, \Psi^{-1}\varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M+1$, which in turn ensures that $x := (\Psi^*)^{-1} y$ satisfies $|\langle x, \varphi_{M;n}\rangle| = b_n$ for every $n = 1, \ldots, M + 1$. □

Based on Theorem 3.9, there is no polynomial-time algorithm to perform phase retrieval for minimal almost injective ensembles, assuming $\mathsf{P} \neq \mathsf{NP}$. On the other hand, there exist ensembles of size $2M - 1$ for which phase retrieval is particularly efficient. For example, letting $\delta_{M;m} \in \mathbb{R}^M$ denote the mth identity basis element, consider the ensemble $\Phi_M := \{\delta_{M;m}\}_{m=1}^M \cup \{\delta_{M;1} + \delta_{M;m}\}_{m=2}^M$; then one can reconstruct (up to global phase) any x whose first entry is nonzero by first taking $\hat{x}(1) :=$

$|\langle x, \delta_{M;1}\rangle|$, and then taking

$$\hat{x}(m) := \frac{1}{2\hat{x}(1)}\Big(|\langle x, \delta_{M;1} + \delta_{M;m}\rangle|^2 - |\langle x, \delta_{M;1}\rangle|^2 - |\langle x, \delta_{M;m}\rangle|^2\Big) \qquad \forall m = 2, \ldots, M.$$

Intuitively, we expect a redundancy threshold that determines whether phase retrieval can be efficient, and this suggests the following open problem: What is the smallest C for which there exists a family of ensembles of size $N = CM + o(M)$ such that phase retrieval can be performed in polynomial time?

We now consider an interesting special case of Problem 3.8 for which an approximate solution can be computed in polynomial time: Take the ensemble $\Phi \in \mathbb{R}^M$ to be the $M \times M$ identity matrix with the all-ones vector as its $(M+1)$st column. Then $|(\Phi^* x)(n)| = b_n$ for all $n = 1, \ldots, M$ (abbreviated $|\Phi^* x| = b$), where

$$b_n := \begin{cases} |x_n| & \text{if } n = 1, \ldots, M \\ \left|\displaystyle\sum_{m=1}^{M} x_m\right| & \text{if } n = M+1. \end{cases}$$

With this notation, we introduce Algorithm 2, which approximately solves the phase retrieval problem for the intensity measurements $|\Phi^* x| = b$. To discuss the performance of this algorithm, it helps to consider the map $\sqrt{\mathcal{A}}$, defined entrywise by $(\sqrt{\mathcal{A}}(x))(n) := |\langle x, \varphi_n\rangle|$. This mapping is actually a near-isometry under a certain metric and, unlike $\mathcal{A}$, it admits desirable performance guarantees (for details, see Section 4.1 of this paper).

Lemma 3.12. *For $M \geq 2$, let Φ be the $M \times (M+1)$ matrix $[I|w]$, $w_n = 1$ for all $n = 1, \ldots, M$, and take $c = \varepsilon/2\sqrt{M}$ for any $\varepsilon > 0$. Then Algorithm 2 produces an estimate $\hat{x}$ such that*

$$\frac{\|\sqrt{\mathcal{A}}(\hat{x}) - \sqrt{\mathcal{A}}(x)\|_2}{\|x\|_2} \leq \varepsilon$$

after $O(M^{2.5}\varepsilon^{-1})$ operations.

Algorithm 2 Approximate $|\Phi^* x| = b$ solver for $\Phi = [I|w]$, $w_n = 1$ for all $n = 1, \dots, M$

```
Input: Rational vector $b$ of length $M + 1$
Output: Approximate solution $\hat{x}$ to $|\Phi^* x| = b$
  Fix a threshold $c$ and set $p := \frac{1}{2} \sum_{m=1}^{M+1} b_m$
  Initialize an $(M + 1) \times M$ matrix $S$ of zeros
  for $j = 1$ to $M$ do
    $S = \begin{bmatrix} s_0 & s_0 + b_j \\ s_1 & s_1 + \delta_j w^T \end{bmatrix}$        {$\delta_j$ is the discrete Dirac-$\delta$ function at $j$}
    $S = \text{sort}\{S\}$        {Sort $S$ by its first row}
    for $k = 1$ to {number of columns in $S$} do
      Remove the $k^{\text{th}}$ column of $S$ if its first entry is greater than $p$ or within $cp/M$
      of the first entry in the $(k-1)^{\text{st}}$ column
    end for
  end for
  if First entry of the last column of $S$ is greater than or equal to $(1 - c)p$ then
    Define $\varepsilon_m = (-1)^{a_m}$ where $(a_1, \dots, a_m)$ are the remaining entries of the last
    column of $S$
    Output: $\hat{x} = (\varepsilon_m b_m)_{m=1}^M$
  else
    Ouput: "INCONSISTENT"
  end if
```

Proof. Suppose $\hat{x} \in \mathbb{R}^M$ is the estimate produced by Algorithm 2. Note that the algorithm guarantees the first M entries of $\sqrt{\mathcal{A}}(\hat{x})$ are identical to those of $\sqrt{\mathcal{A}}(x)$, and so $\|\sqrt{\mathcal{A}}(\hat{x}) - \sqrt{\mathcal{A}}(x)\|_2 = \Big| |\sum_{m=1}^M \hat{x}_m| - |\sum_{m=1}^M x_m| \Big|$. Since

$$p = \frac{1}{2} \sum_{m=1}^{M+1} b_m = \frac{1}{2} \sum_{n=1}^{M+1} |(\Phi^* x)(n)| = \frac{1}{2} \left(\sum_{m=1}^{M} |x_m| + \left| \sum_{m=1}^{M} x_m \right| \right),$$

we see that $|\sum_{m=1}^M x_m| = 2p - \sum_{m=1}^M |x_m| = 2p - \|x\|_1$. Moreover, we have

$$\left| \sum_{m=1}^{M} \hat{x}_m \right| = \left| \sum_{\hat{x}_m \geq 0} |\hat{x}_m| - \sum_{\hat{x}_m < 0} |\hat{x}_m| \right| = \left| \|\hat{x}\|_1 - 2 \sum_{\hat{x}_m < 0} |\hat{x}_m| \right|,$$

and so an application of the triangle inequality yields

$$\begin{aligned}\|\sqrt{\mathcal{A}}(\hat{x})-\sqrt{\mathcal{A}}(x)\|_2 &= \bigg|\Big|\sum_{m=1}^{M}\hat{x}_m\Big|-\Big|\sum_{m=1}^{M}x_m\Big|\bigg| = \bigg|\Big|\|\hat{x}\|_1-2\sum_{\hat{x}_m<0}|\hat{x}_m|\Big|-|2p-\|x\|_1|\bigg|\\ &\leq \bigg|\Big(\|\hat{x}\|_1-2\sum_{\hat{x}_m<0}|\hat{x}_m|\Big)-(2p-\|x\|_1)\bigg| = 2\Big(p-\sum_{\hat{x}_m<0}|\hat{x}_m|\Big).\end{aligned}$$

Since the algorithm ensures that $\sum_{\hat{x}_m<0}|\hat{x}_m|\geq(1-c)p$, this simplifies to

$$\|\sqrt{\mathcal{A}}(\hat{x})-\sqrt{\mathcal{A}}(x)\|_2\leq 2cp\leq 2\sqrt{M}\,c\|x\|_2=\varepsilon\|x\|_2,$$

which rearranges to give the desired bound. To complete the proof, we count operations. The majority of the work in Algorithm 2 is done within the first for loop. In fact, the remainder of the algorithm is performed in $O(M)$ steps, so we will only focus on the first loop. At each iteration, the number of operations performed on the matrix S is dependent on the number of columns since for each column we add a new column by incorporating the next entry of the vector b. Due to the trimming step in the second for loop, however, we limit the number of columns kept at each iteration, thereby limiting the number of operations performed. Since there can be no more than $p/(cp/M))=M/c$ columns at any iteration, an upper bound on the number of operations is $M/c+O(M)=2M^{2.5}/\varepsilon+O(M)=O(M^{2.5}\varepsilon^{-1})$. □

Lemma 3.12 shows that Algorithm 2 produces an estimate whose intensity measurements approximate the true intensity measurements. As such, one can obtain an approximate solution to the phase retrieval problem for the ensemble Φ in polynomial time if willing to work with the estimated intensity measurements:

Theorem 3.13. *For $M\geq 2$, let Φ be the $M\times(M+1)$ matrix $[I|w]$, $w_n=1$ for all $n=1,\ldots,M$, and suppose an estimate $\hat{x}$ for $x\in\mathbb{R}^M$ is produced by Algorithm 2. Then for every nonempty proper subset $S\subseteq\{1,2,\ldots,M+1\}$ and any $\varepsilon>0$, $|\sum_{m\in S}x_m|\leq\frac{1}{2}\varepsilon\|x\|_2$ and $\hat{x}=\pm x$.*

Proof. We seek the contrapositive. First recall that the estimate $\hat{x}$ produced by Algorithm 2 has the property that $\|\sqrt{\mathcal{A}}(\hat{x})-\sqrt{\mathcal{A}}(x)\|_2 = \Big||\sum_{m=1}^{M}\hat{x}_m| - |\sum_{m=1}^{M}x_m|\Big|$. Now suppose that for every nonempty proper subset $S \subseteq \{1,2,\dots,M+1\}$ and any $\varepsilon > 0$ we have $|\sum_{m\in S}x_m| > \frac{1}{2}\varepsilon\|x\|_2$ and assume $\hat{x} \neq \pm x$. Then there exists a nonempty proper subset $S \subseteq \{1,2,\dots,M+1\}$ such that $\hat{x}_m = x_m$ for all $m \in S$ and $\hat{x}_m = -x_m$ for all $m \in S^{\mathrm{c}}$. Thus,

$$\begin{aligned}\|\sqrt{\mathcal{A}}(\hat{x})-\sqrt{\mathcal{A}}(x)\|_2 &= \bigg|\Big|\sum_{m=1}^{M}\hat{x}_m\Big| - \Big|\sum_{m=1}^{M}x_m\Big|\bigg| \\ &= \bigg|\Big|\sum_{m\in S}x_m - \sum_{m\in S^{\mathrm{c}}}x_m\Big| - \Big|\sum_{m\in S}x_m + \sum_{m\in S^{\mathrm{c}}}x_m\Big|\bigg| \\ &= 2\min\left\{\Big|\sum_{m\in S}x_m\Big|, \Big|\sum_{m\in S^{\mathrm{c}}}x_m\Big|\right\},\end{aligned}$$

where the final equality follows from the relation $\big||a|-|b|\big| = \min\{|a+b|,|a-b|\}$ for every $a,b \in \mathbb{R}$. By the assumption on x we then have

$$\|\sqrt{\mathcal{A}}(\hat{x})-\sqrt{\mathcal{A}}(x)\|_2 = 2\min\left\{\Big|\sum_{m\in S}x_m\Big|, \Big|\sum_{m\in S^{\mathrm{c}}}x_m\Big|\right\} > \varepsilon\|x\|_2,$$

violating the result of Lemma 3.12. Hence, $\hat{x}$ could not have been produced by Algorithm 2, as desired. □

To be clear, in the average case it is still highly unlikely that the true signal x is recoverable in non-exponential time from this type of ensemble. However, this isn't too surprising, since it is expected that the smallest constant C for which there exists a family of ensembles of size $N = CM + o(M)$ such that phase retrieval can be performed in polynomial time is greater than one.

IV. The stability of phase retrieval

In order for methods of phase retrieval to be useful for practical applications, they must be able to combat noise. At the very least, we require some semblance of continuity in the intensity measurement mapping; that is, if a signal's intensity measurements are perturbed slightly (e.g., by noise in the measurement process), we seek a bound on the "closeness" of the estimated signal to the true signal. This concept is known as stability, and it is the focus of this chapter. Here, we analyze stability in phase retrieval for both the worst and average cases. For the former, we develop a new condition which strengthens the complement property of Section 2.1; for the latter, we use a stochastic noise model to develop stronger versions of the injectivity characterizations of Chapter II.

4.1 Stability in the worst case

As far as applications are concerned, the stability of reconstruction is perhaps the most important consideration. To date, the only known stability results come from PhaseLift [27], the polarization method [3], and a very recent paper of Eldar and Mendelson [49]. This last paper focuses on the real case, and analyzes how well subgaussian random measurement vectors distinguish signals, thereby yielding some notion of stability which is independent of the reconstruction algorithm used. In particular, given independent random measurement vectors $\{\varphi_n\}_{n=1}^N \subseteq \mathbb{R}^M$, Eldar and Mendelson evaluated measurement separation by finding a constant C such that

$$\|\mathcal{A}(x)-\mathcal{A}(y)\|_1 \geq C\|x-y\|_2\|x+y\|_2 \qquad \forall x,y\in\mathbb{R}^M, \tag{25}$$

where $\mathcal{A}\colon \mathbb{R}^M \to \mathbb{R}^N$ is the intensity measurement process defined by $(\mathcal{A}(x))(n) := |\langle x,\varphi_n\rangle|^2$. With this, we can say that if $\mathcal{A}(x)$ and $\mathcal{A}(y)$ are close, then x must be close to either $\pm y$, and even closer for larger C. By the contrapositive, distant signals will not be confused in the measurement domain because $\mathcal{A}$ does a good job of separating them.

One interesting feature of (25) is that increasing the lengths of the measurement vectors $\{\varphi_n\}_{n=1}^N$ will in turn increase C, meaning the measurements are better separated. As such, for any given magnitude of noise, one can simply amplify the measurement process so as to drown out the noise and ensure stability. However, such amplification could be rather expensive, and so this motivates a different notion of stability—one that is invariant to how the measurement ensemble is scaled. One approach is to build on intuition from Lemma 2.8; that is, a super analysis operator is intuitively more stable if its null space is distant from all rank-2 operators simultaneously; since this null space is invariant to how the measurement vectors are scaled, this is one prospective (and particularly geometric) notion of stability. In this section, we will focus on another alternative. Note that $\mathrm{d}(x,y) := \min\{\|x-y\|,\|x+y\|\}$ defines a metric on $\mathbb{R}^M/\{\pm1\}$, and consider the following:

Definition 4.1. *We say $f\colon \mathbb{R}^M/\{\pm1\} \to \mathbb{R}^N$ is C-stable if for every $\mathrm{SNR} > 0$, there exists an estimator $g\colon \mathbb{R}^N \to \mathbb{R}^M/\{\pm1\}$ such that for every nonzero signal $x \in \mathbb{R}^M/\{\pm1\}$ and adversarial noise term z with $\|z\|^2 \leq \|f(x)\|^2/\mathrm{SNR}$, the relative error in reconstruction satisfies*

$$\frac{\mathrm{d}\big(g(f(x)+z),x\big)}{\|x\|} \leq \frac{C}{\sqrt{\mathrm{SNR}}}.$$

According to this definition, f is more stable when C is smaller. Also, because of the SNR (signal-to-noise ratio) model, f is C-stable if and only if every nonzero multiple of f is also C-stable. Indeed, taking $\tilde{f} := cf$ for some nonzero scalar c, then for every adversarial noise term $\tilde{z}$ which is admissible for $\tilde{f}$ and SNR, we have that $z := \tilde{z}/c$ is admissible for $f(x)$ and SNR; as such, $\tilde{f}$ inherits f's C-stability by using the estimator $\tilde{g}$ defined by $\tilde{g}(y) := g(y/c)$. Overall, this notion of stability offers the invariance to scaling we originally desired. With this, if we find a measurement process f which is C-stable with minimal C, at that point, we can take advantage of noise with bounded magnitude by amplifying f (and thereby effectively increasing SNR) until the relative error in reconstruction is tolerable.

Now that we have a notion of stability, we provide a sufficient condition:

Theorem 4.2. *Suppose f is bilipschitz, that is, there exist constants $0 < \alpha \leq \beta < \infty$ such that*

$$\alpha\,\mathrm{d}(x,y) \leq \|f(x)-f(y)\| \leq \beta\,\mathrm{d}(x,y) \qquad \forall x,y \in \mathbb{R}^M/\{\pm 1\}.$$

If $f(0)=0$, then f is $\frac{2\beta}{\alpha}$-stable.

Proof. Consider the projection function $P\colon \mathbb{R}^N \rightarrow \mathbb{R}^N$ defined by

$$P(y) := \operatorname*{arg\,min}_{y' \in \mathrm{range}(f)} \|y'-y\| \qquad \forall y \in \mathbb{R}^N.$$

In cases where the minimizer is not unique, we will pick one of them to be $P(y)$. For P to be well-defined, we claim it suffices for range(f) to be closed. Indeed, this ensures that a minimizer always exists; since $0 \in$ range(f), any prospective minimizer must be no farther from y than 0 is, meaning we can equivalently minimize over the intersection of range(f) and the closed ball of radius $\|y\|$ centered at y; this intersection is compact, and so a minimizer necessarily exists. In order to avoid using the axiom of choice, we also want a systematic method of breaking ties when the minimizer is not unique, but this can be done using lexicographic ideas provided range(f) is closed.

We now show that range(f) is, in fact, closed. Pick a convergent sequence $\{y_n\}_{n=1}^{\infty} \subseteq$ range(f). This sequence is necessarily Cauchy, which means the corresponding sequence of inverse images $\{x_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^M/\{\pm 1\}$ is also Cauchy (using the lower Lipschitz bound $\alpha > 0$). Arbitrarily pick a representative $z_n \in \mathbb{R}^M$ for each x_n. Then $\{z_n\}_{n=1}^{\infty}$ is bounded, and thus has a subsequence that converges to some $z \in \mathbb{R}^M$. Denote $x := \{\pm z\} \in \mathbb{R}^M/\{\pm 1\}$. Then $\mathrm{d}(x_n, x) \leq \|z_n - z\|$, and so $\{x_n\}_{n=1}^{\infty}$

has a subsequence which converges to x. Since $\{x_n\}_{n=1}^\infty$ is also Cauchy, we therefore have $x_n \to x$. Then the upper Lipschitz bound $\beta < \infty$ gives that $f(x) \in \mathrm{range}(f)$ is the limit of $\{y_n\}_{n=1}^\infty$.

Now that we know P is well-defined, we continue. Since $\alpha > 0$, we know f is injective, and so we can take $g := f^{-1} \circ P$. In fact, α^{-1} is a Lipschitz bound for f^{-1}, implying

$$\mathrm{d}\big(g(f(x)+z),x\big) = \mathrm{d}\Big(f^{-1}\big(P(f(x)+z)\big), f^{-1}\big(f(x)\big)\Big) \leq \alpha^{-1}\|P(f(x)+z) - f(x)\|. \tag{26}$$

Furthermore, the triangle inequality and the definition of P together give

$$\begin{aligned}\|P(f(x)+z) - f(x)\| &\leq \|P(f(x)+z) - (f(x)+z)\| + \|z\| \\ &\leq \|f(x) - (f(x)+z)\| + \|z\| = 2\|z\|. \end{aligned} \tag{27}$$

Combining (26) and (27) then gives

$$\frac{\mathrm{d}\big(g(f(x)+z),x\big)}{\|x\|} \leq 2\alpha^{-1}\frac{\|z\|}{\|x\|} \leq \frac{2\alpha^{-1}}{\sqrt{\mathrm{SNR}}}\frac{\|f(x)\|}{\|x\|} = \frac{2\alpha^{-1}}{\sqrt{\mathrm{SNR}}}\frac{\|f(x)-f(0)\|}{\|x-0\|} \leq \frac{2\beta/\alpha}{\sqrt{\mathrm{SNR}}},$$

as desired. □

Note that the "project-and-invert" estimator we used to demonstrate stability is far from new. For example, if the noise were modeled as Gaussian random, then project-and-invert is precisely the maximum likelihood estimator. However, stochastic noise models warrant a much deeper analysis, since in this regime one is often concerned with the bias and variance of estimates. As such, we will investigate these issues in the next section. Another example of project-and-invert is the Moore-Penrose pseudoinverse of an $N \times M$ matrix A of rank M. Using the obvious reformulation of C-stable in this linear case, it can be shown that C is the condition number of A, meaning α and β are analogous to the smallest and largest singular values. The extra factor of 2 in the stability constant of Theorem 4.2 is an artifact of the nonlinear setting: For the sake of illustration, suppose range(f) is the unit circle and $f(x) = (-1, 0)$ but $z = (1 + \varepsilon, 0)$; then $P(f(x) + z) = (1, 0)$, which is just shy of $2\|z\|$ away from $f(x)$. This sort of behavior is not exhibited in the linear case, in which range(f) is a subspace.

Having established the sufficiency of bilipschitz for stability, we now note that $\mathcal{A}$ is *not* bilipschitz. In fact, more generally, $\mathcal{A}$ fails to satisfy any Hölder condition. To see this, pick some nonzero measurement vector φ_n and scalars $C > 0$ and $\alpha \geq 0$. Then

$$\begin{aligned}\frac{\|\mathcal{A}((C+1)\varphi_n) - \mathcal{A}(\varphi_n)\|}{\mathrm{d}((C+1)\varphi_n, \varphi_n)^\alpha} &= \frac{1}{\|C\varphi_n\|^\alpha}\bigg(\sum_{n'=1}^{N}\Big(|\langle (C+1)\varphi_n, \varphi_{n'}\rangle|^2 - |\langle \varphi_n, \varphi_{n'}\rangle|^2\Big)^2\bigg)^{1/2}\\ &= \frac{(C+1)^2 - 1}{C^\alpha}\frac{\|\mathcal{A}(\varphi_n)\|}{\|\varphi_n\|^\alpha}.\end{aligned}$$

Furthermore, $\|\mathcal{A}(\varphi_n)\| \geq |(\mathcal{A}(\varphi_n))(n)| = \|\varphi_n\|^4 > 0$, while $\frac{(C+1)^2-1}{C^\alpha}$ diverges as $C \to \infty$, assuming $\alpha \leq 1$; when $\alpha > 1$, it also diverges as $C \to 0$, but this case is not interesting for infamous reasons [73].

All is not lost, however. As we will see, with this notion of stability, it happens to be more convenient to consider the map $\sqrt{\mathcal{A}}$, defined entrywise by $(\sqrt{\mathcal{A}}(x))(n) := |\langle x, \varphi_n\rangle|$. Considering Theorem 4.2, we are chiefly interested in the optimal constants $0 < \alpha \leq \beta < \infty$ for which

$$\alpha\,\mathrm{d}(x, y) \leq \|\sqrt{\mathcal{A}}(x) - \sqrt{\mathcal{A}}(y)\| \leq \beta\,\mathrm{d}(x, y) \qquad \forall x, y \in \mathbb{R}^M/\{\pm 1\}. \tag{28}$$

In particular, Theorem 4.2 guarantees more stability when α and β are closer together; this indicates that when suitably scaled, we want $\sqrt{\mathcal{A}}$ to act as a near-isometry, despite being a nonlinear function. The following lemma gives the upper Lipschitz constant:

Lemma 4.3. *The upper Lipschitz constant for $\sqrt{\mathcal{A}}$ is $\beta = \|\Phi^*\|_2$.*

Proof. By the reverse triangle inequality, we have

$$\big||a|-|b|\big| \leq \min\big\{|a-b|,|a+b|\big\} \qquad \forall a,b\in\mathbb{R}.$$

Thus, for all $x,y\in\mathbb{R}^M/\{\pm1\}$,

$$\begin{aligned}\|\sqrt{\mathcal{A}}(x)-\sqrt{\mathcal{A}}(y)\|^2 &= \sum_{n=1}^{N}\big||\langle x,\varphi_n\rangle|-|\langle y,\varphi_n\rangle|\big|^2\\ &\leq \sum_{n=1}^{N}\bigg(\min\Big\{|\langle x-y,\varphi_n\rangle|,|\langle x+y,\varphi_n\rangle|\Big\}\bigg)^2\\ &\leq \min\Big\{\|\Phi^*(x-y)\|^2,\|\Phi^*(x+y)\|^2\Big\}\\ &\leq \|\Phi^*\|_2^2\big(\mathrm{d}(x,y)\big)^2. \qquad (29)\end{aligned}$$

Furthermore, picking a nonzero $x\in\mathbb{R}^M$ such that $\|\Phi^*x\|=\|\Phi^*\|_2\|x\|$ gives

$$\|\sqrt{\mathcal{A}}(x)-\sqrt{\mathcal{A}}(0)\|=\|\sqrt{\mathcal{A}}(x)\|=\|\Phi^*x\|=\|\Phi^*\|_2\|x\|=\|\Phi^*\|_2\,\mathrm{d}(x,0),$$

thereby achieving equality in (29). □

The lower Lipschitz bound is much more difficult to determine. Our approach to analyzing this bound is based on the following definition:

Definition 4.4. *We say an $M\times N$ matrix Φ satisfies the σ-strong complement property (σ-SCP) if*

$$\max\big\{\lambda_{\min}(\Phi_S\Phi_S^*),\lambda_{\min}(\Phi_{S^c}\Phi_{S^c}^*)\big\}\geq\sigma^2$$

for every $S\subseteq\{1,\ldots,N\}$.

This is a numerical version of the complement property discussed earlier (Section 2.1). It bears some resemblance to other matrix properties, namely combinatorial properties regarding the conditioning of submatrices, e.g., the restricted isometry property [23], the Kadison-Singer problem [31] and numerically erasure-robust frames [51]. We are interested in σ-SCP because it is very related to the lower Lipschitz bound in (28):

Theorem 4.5. *The lower Lipschitz constant for $\sqrt{\mathcal{A}}$ satisfies*

$$\sigma \leq \alpha \leq \sqrt{2}\sigma,$$

where σ is the largest scalar for which Φ has the σ-strong complement property.

Proof. By analogy with the proof of Theorem 2.2, we start by proving the upper bound. Pick $\varepsilon > 0$ and note that Φ is not $(\sigma+\varepsilon)$-SCP. Then there exists $S \subseteq \{1,\ldots,N\}$ such that both $\lambda_{\min}(\Phi_S\Phi_S^*) < (\sigma+\varepsilon)^2$ and $\lambda_{\min}(\Phi_{S^c}\Phi_{S^c}^*) < (\sigma+\varepsilon)^2$. This implies that there exist unit (eigen) vectors $u, v \in \mathbb{R}^M$ such that $\|\Phi_S^* u\| < (\sigma+\varepsilon)\|u\|$ and $\|\Phi_{S^c}^* v\| < (\sigma+\varepsilon)\|v\|$. Taking $x := u+v$ and $y := u-v$ then gives

$$\begin{aligned}
\|\sqrt{\mathcal{A}}(x) - \sqrt{\mathcal{A}}(y)\|^2 &= \sum_{n=1}^{N} \big||\langle u+v, \varphi_n\rangle| - |\langle u-v, \varphi_n\rangle|\big|^2 \\
&= \sum_{n\in S} \big||\langle u+v, \varphi_n\rangle| - |\langle u-v, \varphi_n\rangle|\big|^2 \\
&\qquad + \sum_{n\in S^c} \big||\langle u+v, \varphi_n\rangle| - |\langle u-v, \varphi_n\rangle|\big|^2 \\
&\leq 4\sum_{n\in S} |\langle u, \varphi_n\rangle|^2 + 4\sum_{n\in S^c} |\langle v, \varphi_n\rangle|^2,
\end{aligned}$$

where the last step follows from the reverse triangle inequality. Next, we apply our assumptions on u and v:

$$\begin{aligned}
\|\sqrt{\mathcal{A}}(x) - \sqrt{\mathcal{A}}(y)\|^2 &\leq 4\big(\|\Phi_S^* u\|^2 + \|\Phi_{S^c}^* v\|^2\big) \\
&< 4(\sigma+\varepsilon)^2\big(\|u\|^2 + \|v\|^2\big) \\
&= 8(\sigma+\varepsilon)^2 \min\big\{\|u\|^2, \|v\|^2\big\} = 2(\sigma+\varepsilon)^2\big(\mathrm{d}(x,y)\big)^2.
\end{aligned}$$

Thus, $\alpha < \sqrt{2}(\sigma + \varepsilon)$ for all $\varepsilon > 0$, and so $\alpha \leq \sqrt{2}\sigma$.

Next, to prove the lower bound, take $\varepsilon > 0$ and pick $x, y \in \mathbb{R}^M/\{\pm 1\}$ such that

$$(\alpha + \varepsilon)\,\mathrm{d}(x, y) > \|\sqrt{\mathcal{A}}(x) - \sqrt{\mathcal{A}}(y)\|.$$

We will show that Φ is not $(\alpha + \varepsilon)$-SCP. To this end, pick

$$S := \{n : \operatorname{sign}\langle x, \varphi_n\rangle = -\operatorname{sign}\langle y, \varphi_n\rangle\}$$

and define $u := x + y$ and $v := x - y$. Then the definition of S gives

$$\|\Phi_S^* u\|^2 = \sum_{n \in S} |\langle x, \varphi_n\rangle + \langle y, \varphi_n\rangle|^2 = \sum_{n \in S} \big||\langle x, \varphi_n\rangle| - |\langle y, \varphi_n\rangle|\big|^2,$$

and similarly $\|\Phi_{S^c}^* v\|^2 = \sum_{n \in S^c} \big||\langle x, \varphi_n\rangle| - |\langle y, \varphi_n\rangle|\big|^2$. Adding these together then gives

$$\begin{aligned}\|\Phi_S^* u\|^2 + \|\Phi_{S^c}^* v\|^2 &= \sum_{n=1}^{N} \big||\langle x, \varphi_n\rangle| - |\langle y, \varphi_n\rangle|\big|^2 \\ &= \|\sqrt{\mathcal{A}}(x) - \sqrt{\mathcal{A}}(y)\|^2 < (\alpha + \varepsilon)^2 \big(\mathrm{d}(x, y)\big)^2,\end{aligned}$$

implying both $\|\Phi_S^* u\| < (\alpha + \varepsilon)\|u\|$ and $\|\Phi_{S^c}^* v\| < (\alpha + \varepsilon)\|v\|$. Therefore, Φ is not $(\alpha + \varepsilon)$-SCP, i.e., $\sigma < \alpha + \varepsilon$ for all $\varepsilon > 0$, which in turn implies the desired lower bound. □

Note that all of this analysis specifically treats the real case; indeed, the metric we use would not be appropriate in the complex case. However, just like the complement property is necessary for injectivity in the complex case (Theorem 2.6), it is suspected that the strong complement property is necessary for stability in the complex case, but we have no proof of this.

As an example of how to apply Theorem 4.5, pick M and N to both be even and let $F = \{f_n\}_{n\in\mathbb{Z}_N}$ be the $\frac{M}{2} \times N$ matrix obtained by collecting the first $\frac{M}{2}$ rows of the $N \times N$ discrete Fourier transform matrix with entries of unit-modulus. Next, take $\Phi = \{\varphi_n\}_{n\in\mathbb{Z}_N}$ to be the $M \times N$ matrix formed by stacking the real and imaginary parts of F and normalizing the resulting columns (i.e., multiplying by $\sqrt{2/M}$). Then Φ happens to be a *self-localized finite frame* due to the rapid decay in coherence between columns. To be explicit, first note that

$$\begin{aligned}
|\langle\varphi_n,\varphi_{n'}\rangle|^2 &= \tfrac{4}{M^2}|\langle\operatorname{Re} f_n,\operatorname{Re} f_{n'}\rangle + \langle\operatorname{Im} f_n,\operatorname{Im} f_{n'}\rangle|^2\\
&\leq \tfrac{4}{M^2}\bigg|\Big(\langle\operatorname{Re} f_n,\operatorname{Re} f_{n'}\rangle + \langle\operatorname{Im} f_n,\operatorname{Im} f_{n'}\rangle\Big)\\
&\qquad\qquad + \mathrm{i}\Big(\langle\operatorname{Im} f_n,\operatorname{Re} f_{n'}\rangle - \langle\operatorname{Re} f_n,\operatorname{Im} f_{n'}\rangle\Big)\bigg|^2\\
&= \tfrac{4}{M^2}|\langle f_n,f_{n'}\rangle|^2,
\end{aligned}$$

and furthermore, when $n \neq n'$, the geometric sum formula gives

$$|\langle f_n,f_{n'}\rangle|^2 = \bigg|\sum_{m=0}^{M-1} e^{2\pi \mathrm{i} m(n-n')/N}\bigg|^2 = \frac{\sin^2(M\pi(n-n')/N)}{\sin^2(\pi(n-n')/N)} \leq \frac{1}{\sin^2(\pi(n-n')/N)}.$$

Taking $u := \varphi_0$, $v := \varphi_{N/2}$ and $S := \{n : \frac{N}{4} \leq n < \frac{3N}{4}\}$, we then have

$$\frac{\|\Phi_S^* u\|^2}{\|u\|^2} = \|\Phi_S^* u\|^2 = \sum_{n\in S}|\langle\varphi_0,\varphi_n\rangle|^2 \leq \frac{4}{M^2}\sum_{n\in S}\frac{1}{\sin^2(\pi n/N)} \leq \frac{4}{M^2}\cdot\frac{N/2}{\sin^2(\pi/4)} = \frac{4N}{M^2},$$

and similarly for $\frac{\|\Phi_{S^c}^* v\|^2}{\|v\|^2}$. As such, if $N = o(M^2)$, then Φ is σ-SCP only if σ vanishes, meaning phase retrieval with Φ necessarily lacks the stability guarantee of Theorem 4.5. As a rule of thumb, self-localized frames fail to provide stable phase retrieval for this very reason; just as we cannot stably distinguish between $\varphi_0 + \varphi_{N/2}$ and $\varphi_0 - \varphi_{N/2}$ in this case, in general, signals consisting of "distant" components bring similar instability. This intuition was first pointed out by Irene Waldspurger—here it is simply made more rigorous with the notion of σ-SCP. This means that stable phase retrieval from localized measurements must either use prior information about the signal (e.g., connected support) or additional measurements; indeed, this dichotomy has already made its mark on the Fourier-based phase retrieval literature [50, 65].

We can also apply the strong complement property to show that certain (random) ensembles produce stable measurements. We will use the following lemma, which is proved in the proof of Lemma 4.1 in [35]:

Lemma 4.6. *Given $n \geq m \geq 2$, draw a real $m \times n$ matrix G of independent standard normal entries. Then*

$$\Pr\left(\lambda_{\min}(GG^*) \leq \frac{n}{t^2}\right) \leq \frac{1}{\Gamma(n-m+2)}\left(\frac{n}{t}\right)^{n-m+1} \qquad \forall t > 0.$$

Theorem 4.7. *Draw an $M \times N$ matrix Φ with independent standard normal entries, and denote $R = \frac{N}{M}$. Provided $R > 2$, then for every $\varepsilon > 0$, Φ has the σ-strong complement property with*

$$\sigma = \frac{1}{\sqrt{2}e^{1+\varepsilon/(R-2)}} \cdot \frac{N-2M+2}{2^{R/(R-2)}\sqrt{N}},$$

with probability greater than or equal to $1 - e^{-\varepsilon M}$.

Proof. Fix M and N, and consider the function $f\colon (M-2, \infty) \to (0, \infty)$ defined by

$$f(x) := \frac{1}{\Gamma(x-M+2)}(\sigma\sqrt{x})^{x-M+1}.$$

To simplify the analysis, we will assume that N is even, but the proof can be amended to account for the odd case. Applying Lemma 4.6, we have for every subset $S \subseteq \{1, \ldots, N\}$ of size K that $\Pr(\lambda_{\min}(\Phi_S\Phi_S^*) < \sigma^2) \leq f(K)$, provided $K \geq M$, and similarly $\Pr(\lambda_{\min}(\Phi_{S^c}\Phi_{S^c}^*) < \sigma^2) \leq f(N-K)$, provided $N-K \geq M$. We will use this to bound the probability that Φ is not σ-SCP. Since $\lambda_{\min}(\Phi_{S^c}\Phi_{S^c}^*) = 0$ whenever

$|S| \geq N-M+1$ and $\lambda_{\min}(\Phi_S\Phi_S^*) \leq \lambda_{\min}(\Phi_T\Phi_T^*)$ whenever $S \subseteq T$, a union bound gives

$$\begin{aligned}
&\Pr\Big(\Phi \text{ is not } \sigma\text{-SCP}\Big)\\
&\quad= \Pr\Big(\exists S \subseteq \{1,\ldots,N\} \text{ s.t. } \lambda_{\min}(\Phi_S\Phi_S^*) < \sigma^2 \text{ and } \lambda_{\min}(\Phi_{S^\mathrm{c}}\Phi_{S^\mathrm{c}}^*) < \sigma^2\Big)\\
&\quad\leq \Pr\Big(\exists S \subseteq \{1,\ldots,N\}, |S| = N-M+1, \text{ s.t. } \lambda_{\min}(\Phi_S\Phi_S^*) < \sigma^2\Big)\\
&\qquad\quad + \Pr\Big(\exists S \subseteq \{1,\ldots,N\}, M \leq |S| \leq N-M,\\
&\qquad\qquad\qquad \text{s.t. } \lambda_{\min}(\Phi_S\Phi_S^*) < \sigma^2 \text{ and } \lambda_{\min}(\Phi_{S^\mathrm{c}}\Phi_{S^\mathrm{c}}^*) < \sigma^2\Big)\\
&\quad\leq \binom{N}{N-M+1} f(N-M+1) + \frac{1}{2}\sum_{K=M}^{N-M}\binom{N}{K} f(K)f(N-K), \qquad (30)
\end{aligned}$$

where the last inequality follows in part from the fact that both $\lambda_{\min}(\Phi_S\Phi_S^*)$ and $\lambda_{\min}(\Phi_{S^\mathrm{c}}\Phi_{S^\mathrm{c}}^*)$ are independent random variables, and the factor $\frac{1}{2}$ is an artifact of double counting partitions. We will further bound each term in (30) to get a simpler expression. First, $\binom{2k}{k} \geq 2^k$ for all k and so

$$\begin{aligned}
f(N-M+1) &\leq \frac{1}{\Gamma(N-2M+3)}(\sigma\sqrt{N})^{N-2M+2}\\
&\leq \frac{1}{\Gamma(N-2M+3)}(\sigma\sqrt{N})^{N-2M+2} \cdot \frac{1}{2^{\frac{N}{2}-M+1}}\binom{N-2M+2}{\frac{N}{2}-M+1}\\
&= f(\tfrac{N}{2})^2.
\end{aligned}$$

Next, we will find that $g(x) := f(x)f(N-x)$ is maximized at $x = \frac{N}{2}$. To do this, we first find the critical points of g. Since $0 = g'(x) = f'(x)f(N-x) - f(x)f'(N-x)$, we have

$$\frac{d}{dy}\log f(y)\bigg|_{y=x} = \frac{f'(x)}{f(x)} = \frac{f'(N-x)}{f(N-x)} = \frac{d}{dy}\log f(y)\bigg|_{y=N-x}. \qquad (31)$$

To analyze this further, we take another derivative:

$$\frac{d^2}{dy^2}\log f(y) = \frac{1}{2y} + \frac{M-1}{2y^2} - \frac{d^2}{dy^2}\log\Gamma(y-M+2). \qquad (32)$$

It is straightforward to see that

$$\begin{aligned}\frac{1}{2y}+\frac{M-1}{2y^2}\leq\frac{1}{y-M+2}&=\int_{y-M+2}^{\infty}\frac{dt}{t^2}\\&<\sum_{k=0}^{\infty}\frac{1}{(y-M+2+k)^2}=\frac{d^2}{dy^2}\log\Gamma(y-M+2),\end{aligned}$$

where the last step uses a series expression for the trigamma function $\psi_1(z) := \frac{d^2}{dz^2}\log\Gamma(z)$; see Section 6.4 of [1]. Applying this to (32) then gives that $\frac{d^2}{dy^2}\log f(y) < 0$, which in turn implies that $\frac{d}{dy}\log f(y)$ is strictly decreasing in y. Thus, (31) requires $x = N - x$, and so $x = \frac{N}{2}$ is the only critical point of g. Furthermore, to see that this is a maximizer, notice that

$$\begin{aligned}g''(\tfrac{N}{2})&=2f(\tfrac{N}{2})^2\cdot\frac{f''(\frac{N}{2})f(\frac{N}{2})-f'(\frac{N}{2})^2}{f(\frac{N}{2})^2}\\&=2f(\tfrac{N}{2})^2\cdot\frac{d}{dy}\frac{f'(y)}{f(y)}\bigg|_{y=\frac{N}{2}}=2f(\tfrac{N}{2})^2\cdot\frac{d^2}{dy^2}\log f(y)\bigg|_{y=\frac{N}{2}}<0.\end{aligned}$$

To summarize, we have that $f(N-M+1)$ and $f(K)f(N-K)$ are both at most $f(\frac{N}{2})^2$. This leads to the following bound on (30):

$$\begin{aligned}\Pr\Big(\Phi\text{ is not }\sigma\text{-SCP}\Big)&\leq\frac{1}{2}\sum_{K=0}^{N}\binom{N}{K}f(\tfrac{N}{2})^2\\&=2^{N-1}f(\tfrac{N}{2})^2=\frac{2^{N-1}}{\Gamma(\frac{N}{2}-M+2)^2}\Big(\sigma\sqrt{\tfrac{N}{2}}\Big)^{N-2M+2}.\end{aligned}$$

Finally, applying the fact that $\Gamma(k+1)\geq e(\frac{k}{e})^k$ gives

$$\begin{aligned}\Pr\Big(\Phi\text{ is not }\sigma\text{-SCP}\Big)&\leq\frac{2^{N-1}}{e^2}\Big(\sigma e\sqrt{2}\cdot\frac{\sqrt{N}}{N-2M+2}\Big)^{N-2M+2}\\&=\frac{2^{RM}}{2e^2}\Big(e^{-\varepsilon/(R-2)}2^{-R/(R-2)}\Big)^{(R-2)M+2}\\&\leq 2^{RM}(e^{\varepsilon}2^R)^{-M}=e^{-\varepsilon M},\end{aligned}$$

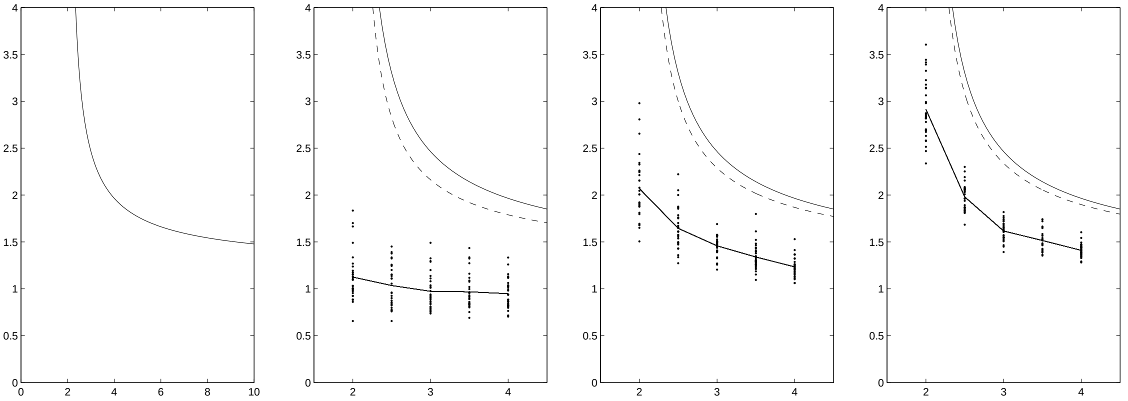


Figure 3: The graph on the left depicts $\log_{10} b(R)$ as a function of R, which is defined in (33). Modulo ε terms, this serves as an upper bound on $\log_{10}(2\|\Phi^*\|_2/\sigma)$ with high probability as $M \to \infty$, where Φ is an $M \times RM$ matrix of independent standard Gaussian entries. Based on Theorem 4.2 (along with Lemma 4.3 and Theorem 4.5), this provides a stability guarantee for the corresponding measurement process, namely $\sqrt{\mathcal{A}}$. Since $\log_{10} b(R)$ exhibits an asymptote at $R = 2$, this gives no stability guarantee for measurement ensembles of redundancy 2. The next three graphs consider the special cases where $M = 2, 4, 6$, respectively. In each case, the dashed curve depicts the slightly stronger upper bound of $\log_{10} a(R, M)$, defined in (33). Also depicted, for each $R \in \{2, 2.5, 3, 3.5, 4\}$, are 30 realizations of $\log_{10}(2\|\Phi^*\|_2/\sigma)$; we provide a piecewise linear graph connecting the sample averages for clarity. Notice that as M increases, $\log_{10} a(R, M)$ approaches $\log_{10} b(R)$; this is easily seen by their definitions in (33). More interestingly, the random realizations also appear to be approaching $\log_{10} b(R)$; this is most notable with the realizations corresponding to $R = 2$. To be clear, we use σ as a proxy for α (see Theorem 4.5) because α is particularly difficult to obtain; as such, we do not plot realizations of $\log_{10}(2\beta/\alpha)$.

as claimed. □

Considering $\|\Phi^*\|_2 \le (1+\varepsilon)(\sqrt{N}+\sqrt{M})$ with probability $\ge 1-2e^{-\varepsilon(\sqrt{N}^2+\sqrt{M})^2/2}$ (see Theorem II.13 of [42]), we can leverage Theorem 4.7 to determine the stability of a Gaussian measurement ensemble. Specifically, by Theorem 4.2 (along with Lemma 4.3 and Theorem 4.5) we have that such measurements are C-stable with

$$C = \frac{2\beta}{\alpha} \le \frac{2\|\Phi^*\|_2}{\sigma} \sim \underbrace{2(\sqrt{N}+\sqrt{M}) \cdot \sqrt{2}e \cdot \frac{2^{R/(R-2)}\sqrt{N}}{N-2M+2}}_{a(R,M)}$$

$$\le \underbrace{2\sqrt{2}e\left(\frac{R+\sqrt{R}}{R-2}\right)2^{R/(R-2)}}_{b(R)} \qquad (33)$$

Figure 3 illustrates these bounds along with different realizations of $2\|\Phi^*\|_2/\sigma$. This suggests that the redundancy of the measurement process is the main factor that determines stability of a random measurement ensemble (and that bounded redundancies suffice for stability). Furthermore, the project-and-invert estimator will yield particularly stable signal reconstruction, although it is not obvious how to efficiently implement this estimator; this is one advantage given by the reconstruction algorithms in [3, 27].

4.2 Stability in the average case

Suppose a random variable Y is drawn according to some unknown member of a parameterized family of probability density functions $\{f(\cdot;\theta)\}_{\theta\in\Omega}$. The Fisher information $J(\theta)$ quantifies how much information about the unknown parameter θ is given by the random variable on average. This is particularly useful in statistical signal processing, where a signal measurement is corrupted by random noise, and the original signal is viewed as a parameter of the random measurement's unknown probability density function; as such, the Fisher information quantifies how useful the noisy measurement is for signal estimation.

In this section, we will apply the theory of Fisher information to evaluate the stability of the intensity measurement mapping $\mathcal{A}$. To do this, we consider a stochastic noise model, that is, given some signal x, we take measurements of the form $Y = \mathcal{A}(x) + Z$, where the entries of Z are independent Gaussian random variables with mean 0 and variance σ^2. We want to use Y to estimate x up to a global phase factor; to simplify the analysis, we will estimate a particular $\theta(x) \equiv x$, specifically (and arbitrarily) x divided by the phase of its last nonzero entry. As such, Y is a random vector with probability density function

$$f(y;\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\|y-\mathcal{A}(\theta)\|^2/2\sigma^2} \qquad \forall y \in \mathbb{R}^N. \tag{34}$$

To be clear, many of the results that follow are consequences of the fact that (34) is a member of the exponential family of distributions; we go through the analysis here since the relevant literature may be less familiar to the phase retrieval community.

With the probability density function (34), we can calculate the Fisher information matrix, defined entrywise by

$$\big(J(\theta)\big)_{ij} := \mathbb{E}\bigg[\bigg(\frac{\partial}{\partial\theta_i}\log f(Y;\theta)\bigg)\bigg(\frac{\partial}{\partial\theta_j}\log f(Y;\theta)\bigg)\bigg|\theta\bigg]. \tag{35}$$

In particular, we have

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\log f(y;\theta) &= \frac{\partial}{\partial\theta_i}\bigg(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\Big(y_n-\big(\mathcal{A}(\theta)\big)_n\Big)^2\bigg)\\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\Big(y_n-\big(\mathcal{A}(\theta)\big)_n\Big)\frac{\partial}{\partial\theta_i}\big(\mathcal{A}(\theta)\big)_n,\end{aligned}$$

and so applying (35) along with the independence of the entries of Z gives

$$\begin{aligned}\big(J(\theta)\big)_{ij} &= \frac{1}{\sigma^4}\sum_{n=1}^{N}\sum_{n'=1}^{N}\frac{\partial}{\partial\theta_i}\big(\mathcal{A}(\theta)\big)_n\frac{\partial}{\partial\theta_j}\big(\mathcal{A}(\theta)\big)_{n'}\mathbb{E}[Z_nZ_{n'}]\\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}\frac{\partial}{\partial\theta_i}\big(\mathcal{A}(\theta)\big)_n\frac{\partial}{\partial\theta_j}\big(\mathcal{A}(\theta)\big)_n.\end{aligned}$$

It remains to take partial derivatives of $\mathcal{A}(\theta)$, but this calculation depends on whether θ is real or complex. In the real case, we have

$$\frac{\partial}{\partial\theta_i}\big(\mathcal{A}(\theta)\big)_n = \frac{\partial}{\partial\theta_i}\bigg(\sum_{m=1}^{M}\theta_m\varphi_n(m)\bigg)^2 = 2\bigg(\sum_{m=1}^{M}\theta_m\varphi_n(m)\bigg)\varphi_n(i).$$

Thus, if we take $\Psi(\theta)$ to be the $M\times N$ matrix whose nth column is $\langle\theta,\varphi_n\rangle\varphi_n$, then the Fisher information matrix can be expressed as $J(\theta)=\frac{4}{\sigma^2}\Psi(\theta)\Psi(\theta)^*$. Interestingly, Theorem 2.2 implies that $J(\theta)$ is necessarily positive definite when $\mathcal{A}$ is injective. To see this, suppose there exists $\theta\in\Omega$ such that $J(\theta)$ has a nontrivial null space. Then $\{\langle\theta,\varphi_n\rangle\varphi_n\}_{n=1}^N$ does not span $\mathbb{R}^M$, and so $S=\{n:\langle\theta,\varphi_n\rangle=0\}$ breaks the complement property. As the following result shows, when $\mathcal{A}$ is injective, the conditioning of $J(\theta)$ lends some insight into stability:

Theorem 4.8. *For $x \in \mathbb{R}^M$, let $Y = \mathcal{A}(x) + Z$ denote noisy intensity measurements with Z having independent $\mathcal{N}(0, \sigma^2)$ entries. Furthermore, define the parameter θ to be x divided by the sign of its last nonzero entry; let $\Omega \subseteq \mathbb{R}^M$ denote all such θ. Then for any unbiased estimator $\hat{\theta}(Y)$ of θ in Ω with a finite $M \times M$ covariance matrix $C(\hat{\theta})$, we have $C(\hat{\theta}) - J(\theta)^{-1}$ is positive semidefinite whenever $\theta \in \operatorname{int}(\Omega)$.*

This result was first given by Balan (see Theorem 4.1 in [5]). Note that the requirement that θ be in the interior of Ω can be weakened to $\theta \neq 0$ by recognizing that our choice for θ (dividing by the sign of the last nonzero entry) was arbitrary. To interpret this theorem, note that

$$\begin{aligned}
\operatorname{Tr}[C(\hat{\theta})] &= \operatorname{Tr}[\mathbb{E}[(\hat{\theta}(Y) - \theta)(\hat{\theta}(Y) - \theta)^{\mathrm{T}}]] \\
&= \mathbb{E}[\operatorname{Tr}[(\hat{\theta}(Y) - \theta)(\hat{\theta}(Y) - \theta)^{\mathrm{T}}]] \\
&= \mathbb{E}[\operatorname{Tr}[(\hat{\theta}(Y) - \theta)^{\mathrm{T}}(\hat{\theta}(Y) - \theta)]] = \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2,
\end{aligned}$$

and so Theorem 4.8 and the linearity of the trace together give $\mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 = \operatorname{Tr}[C(\hat{\theta})] \geq \operatorname{Tr}[J(\theta)^{-1}]$. In the previous section, Definition 4.1 provided a notion of worst-case stability based on the existence of an estimator with small error. By analogy, Theorem 4.8 demonstrates a converse of sorts: that no unbiased estimator will have mean squared error smaller than $\operatorname{Tr}[J(\theta)^{-1}]$. As such, a stable measurement ensemble might minimize $\sup_{\theta \in \Omega} \operatorname{Tr}[J(\theta)^{-1}]$, although this is a particularly cumbersome objective function to work with. More interestingly, Theorem 4.8 provides another numerical strengthening of the complement property (analogous to the σ-strong complement property of the previous section). Unfortunately, we cannot make a more rigorous comparison between the worst- and average-case analyses of stability; indeed, our worst-case analysis exploited the fact that $\sqrt{\mathcal{A}}$ is bilipschitz (which $\mathcal{A}$ is not), and as we shall see, the average-case analysis depends on $\mathcal{A}$ being differentiable (which $\sqrt{\mathcal{A}}$ is not).

To calculate the information matrix in the complex case, we first express our parameter vector in real coordinates: $\theta = (\theta_1 + \mathrm{i}\theta_{M+1}, \theta_2 + \mathrm{i}\theta_{M+2}, \dots, \theta_M + \mathrm{i}\theta_{2M})$, that is, we view θ as a $2M$-dimensional real vector by concatenating its real and imaginary parts. Next, for any arbitrary function $g\colon \mathbb{R}^{2M} \to \mathbb{C}$, the product rule gives

$$\frac{\partial}{\partial\theta_i}|g(\theta)|^2 = \frac{\partial}{\partial\theta_i}g(\theta)\overline{g(\theta)} = \bigg(\frac{\partial}{\partial\theta_i}g(\theta)\bigg)\overline{g(\theta)} + g(\theta)\overline{\bigg(\frac{\partial}{\partial\theta_i}g(\theta)\bigg)} = 2\operatorname{Re} g(\theta)\frac{\partial}{\partial\theta_i}\overline{g(\theta)}. \tag{36}$$

Since we care about partial derivatives of $\mathcal{A}(\theta)$, we take

$$g(\theta) = \langle\theta, \varphi_n\rangle = \sum_{m=1}^{M}(\theta_m + \mathrm{i}\theta_{M+m})\overline{\varphi_n(m)},$$

and so

$$\frac{\partial}{\partial\theta_i}\overline{g(\theta)} = \begin{cases} \varphi_n(i) & \text{if } i \leq M \\ -\mathrm{i}\varphi_n(i-M) & \text{if } i > M. \end{cases} \tag{37}$$

Combining (36) and (37) then gives the following expression for the Fisher information matrix: Take $\Psi(\theta)$ to be the $2M \times N$ matrix whose nth column is formed by stacking the real and imaginary parts of $\langle\theta, \varphi_n\rangle\varphi_n$; then $J(\theta) = \frac{4}{\sigma^2}\Psi(\theta)\Psi(\theta)^*$.

Lemma 4.9. *Take $\widetilde{J}(\theta)$ to be the $(2M-1) \times (2M-1)$ matrix that comes from removing the last row and column of $J(\theta)$. If $\mathcal{A}$ is injective, then $\widetilde{J}(\theta)$ is positive definite for every $\theta \in \operatorname{int}(\Omega)$.*

Proof. First, we note that $J(\theta) = \frac{4}{\sigma^2}\Psi(\theta)\Psi(\theta)^*$ is necessarily positive semidefinite, and so

$$\inf_{\|x\|=1} x^{\mathrm{T}}\widetilde{J}(\theta)x = \inf_{\|x\|=1}[x;0]^{\mathrm{T}}J(\theta)[x;0] \geq \inf_{\|y\|=1} y^{\mathrm{T}}J(\theta)y \geq 0.$$

As such, it suffices to show that $\widetilde{J}(\theta)$ is invertible.

To this end, take any vector x in the null space of $\widetilde{J}(\theta)$. Then defining $y := [x;0] \in \mathbb{R}^{2M}$, we have that $J(\theta)y$ is zero in all but (possibly) the $2M$th entry. As such, $0 = \langle y, J(\theta)y\rangle = \|\frac{2}{\sigma}\Psi(\theta)^*y\|^2$, meaning y is orthogonal to the columns of $\Psi(\theta)$. Since $\mathcal{A}$ is injective, Theorem 2.3 then gives that $y = \alpha \mathrm{i}\theta$ for some $\alpha \in \mathbb{R}$. But since $\theta \in \operatorname{int}(\Omega)$, we have $\theta_M > 0$, and so the $2M$th entry of $\mathrm{i}\theta$ is necessarily nonzero. This means $\alpha = 0$, and so y (and thus x) is trivial. □

Theorem 4.10. *For $x \in \mathbb{C}^M$, let $Y = \mathcal{A}(x)+Z$ denote noisy intensity measurements with Z having independent $\mathcal{N}(0,\sigma^2)$ entries. Furthermore, define the parameter θ to be x divided by the phase of its last nonzero entry, and view θ as a vector in $\mathbb{R}^{2M}$ by concatenating its real and imaginary parts; let $\Omega \subseteq \mathbb{R}^{2M}$ denote all such θ. Then for any unbiased estimator $\hat{\theta}(Y)$ of θ in Ω with a finite $2M\times 2M$ covariance matrix $C(\hat{\theta})$, the last row and column of $C(\hat{\theta})$ are both zero, and the remaining $(2M-1)\times(2M-1)$ submatrix $\widetilde{C}(\hat{\theta})$ has the property that $\widetilde{C}(\hat{\theta}) - \widetilde{J}(\theta)^{-1}$ is positive semidefinite whenever $\theta \in \operatorname{int}(\Omega)$.*

Proof. We start by following the usual proof of the vector parameter Cramer-Rao lower bound (see for example Appendix 3B of [69]). Note that for any $i, j \in \{1,\ldots,2M\}$,

$$\begin{aligned}\int_{\mathbb{R}^N} \big((\hat{\theta}(y))_j - \theta_j\big)\frac{\partial \log f(y;\theta)}{\partial\theta_i} f(y;\theta)dy \qquad\qquad\qquad&\\ = \int_{\mathbb{R}^N} (\hat{\theta}(y))_j \frac{\partial f(y;\theta)}{\partial\theta_i}dy - \theta_j\int_{\mathbb{R}^N}\frac{\partial f(y;\theta)}{\partial\theta_i}dy&\\ = \frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^N}(\hat{\theta}(y))_j f(y;\theta)dy - \theta_j\frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^N} f(y;\theta)dy,&\end{aligned}$$

where the second equality is by differentiation under the integral sign (see Lemma A.1 in Appendix A for details; here, we use the fact that $\hat{\theta}$ has a finite covariance matrix so that $\hat{\theta}_j$ has a finite second moment). Next, we use the facts that $\hat{\theta}$ is unbiased and $f(\cdot;\theta)$ is a probability density function (regardless of θ) to get

$$\int_{\mathbb{R}^N} \big((\hat{\theta}(y))_j - \theta_j\big)\frac{\partial \log f(y;\theta)}{\partial\theta_i} f(y;\theta)dy = \frac{\partial\theta_j}{\partial\theta_i} = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i\neq j\end{cases}.$$

Thus, letting $\nabla_\theta \log f(y;\theta)$ denote the column vector whose ith entry is $\frac{\partial \log f(y;\theta)}{\partial \theta_i}$, we have

$$I = \int_{\mathbb{R}^N} \big(\hat{\theta}(y) - \theta\big)\big(\nabla_\theta \log f(y;\theta)\big)^{\mathrm{T}} f(y;\theta)dy.$$

Equivalently, we have that for all column vectors $a, b \in \mathbb{R}^{2M}$,

$$a^{\mathrm{T}} b = \int_{\mathbb{R}^N} a^{\mathrm{T}} \big(\hat{\theta}(y) - \theta\big)\big(\nabla_\theta \log f(y;\theta)\big)^{\mathrm{T}} b \ f(y;\theta)dy.$$

Next, we apply the Cauchy-Schwarz inequality in f-weighted L^2 space to get

$$\begin{aligned}
\big(a^{\mathrm{T}} b\big)^2 &= \bigg(\int_{\mathbb{R}^N} a^{\mathrm{T}} \big(\hat{\theta}(y) - \theta\big)\big(\nabla_\theta \log f(y;\theta)\big)^{\mathrm{T}} b \ f(y;\theta)dy\bigg)^2 \\
&\le \int_{\mathbb{R}^N} a^{\mathrm{T}} \big(\hat{\theta}(y) - \theta\big)\big(\hat{\theta}(y) - \theta\big)^{\mathrm{T}} a \ f(y;\theta)dy \\
&\qquad\qquad \times \int_{\mathbb{R}^N} b^{\mathrm{T}} \big(\nabla_\theta \log f(y;\theta)\big)\big(\nabla_\theta \log f(y;\theta)\big)^{\mathrm{T}} b \ f(y;\theta)dy \\
&= \big(a^{\mathrm{T}} C(\hat{\theta}) a\big)\big(b^{\mathrm{T}} J(\theta) b\big),
\end{aligned}$$

where the last step follows from pulling vectors out of integrals. Taking $b := [\widetilde{J}(\theta)^{-1}\tilde{a}; 0]$, where $\tilde{a}$ is the first $2M-1$ entries of a, this then implies

$$\big(\tilde{a}^{\mathrm{T}} \widetilde{J}(\theta)^{-1}\tilde{a}\big)^2 = \big(a^{\mathrm{T}} b\big)^2 \le \big(a^{\mathrm{T}} C(\hat{\theta}) a\big)\big(b^{\mathrm{T}} J(\theta) b\big) = \big(a^{\mathrm{T}} C(\hat{\theta}) a\big)\big(\tilde{a}^{\mathrm{T}} \widetilde{J}(\theta)^{-1}\tilde{a}\big). \quad (38)$$

At this point, we note that since the last (complex) entry of $\theta \in \Omega$ is necessarily positive, then as a $2M$-dimensional real vector, the last entry is necessarily zero, and furthermore every unbiased estimator $\hat{\theta}$ in Ω will also vanish in the last entry. It follows that the last row and column of $C(\hat{\theta})$ are both zero. Furthermore, since $\widetilde{J}(\theta)^{-1}$ is positive definite by Lemma 4.9, division in (38) gives

$$\big(\tilde{a}^{\mathrm{T}} \widetilde{J}(\theta)^{-1}\tilde{a}\big) \le \big(a^{\mathrm{T}} C(\hat{\theta}) a\big) = \big(\tilde{a}^{\mathrm{T}} \widetilde{C}(\hat{\theta})\tilde{a}\big),$$

from which the result follows. □

V. The phase error problem in synthetic aperture radar

Now that we've developed an intuition for phase retrieval, we return to the phase error problem in synthetic aperture radar. We begin by formally deriving phase errors in the bistatic setting, at which point we relate the problem to a certain interferometric approach to phase retrieval [3]. This motivates the use of graphs to organize the given SAR data, and then we can leverage an algorithm known as angular synchronization to recover the phase errors. The remainder of the chapter is dedicated to developing this algorithmic approach to solving the phase error problem, and we conclude with simulations that illustrate its stability to noise.

5.1 Synthetic aperture radar

As discussed in Chapter I, SAR is a form of microwave radar that uses relative motion between a source and scene to reconstruct an image of the scene with finer resolution than possible with traditional radar. SAR is typically implemented in a monostatic setting, usually as a single source on a moving platform (e.g., an aircraft or satellite) which repeatedly transmits a fixed microwave signal to the target scene and records the resultant reflected signal. As a result, the scene is imaged at various locations; each reflected signal provides information about the scene from a different perspective, allowing for finer resolution in the final reconstruction. In this section, we will consider airborne SAR, where the radar source is located on a moving aircraft. Since the speed of light far exceeds that of the aircraft, the transmitted and reflected signals will effectively travel to and from the aircraft along the same path.

To see how airborne SAR works, consider a two-dimensional scene

$$D := \{x = (x_1, x_2)^\top \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq \beta^2\},$$

with *magnetic reflectivity* described by the function $\rho\colon \mathbb{R}^2 \to \mathbb{R}$; the reflectivity is taken to be zero outside of the region D for simplicity. The assumption that the scene is two-dimensional is possible only if elevations within the scene are relatively constant in comparison to the radius β; here, we operate under this assumption. Taking the position of the aircraft to be $r \in \mathbb{R}^2$, where $\|r\| \gg \beta$, the Pythagorean Theorem yields

$$\|x - r\|^2 = \big|\langle x - r, \tfrac{r}{\|r\|}\rangle\big|^2 + \big|\langle x - r, \tfrac{Ar}{\|r\|}\rangle\big|^2$$

for every vector $x \in D$, where $A \in \mathbb{R}^{2\times 2}$ is the rotation matrix

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Since r and Ar are orthogonal, we then obtain

$$\big|\langle x - r, \tfrac{Ar}{\|r\|}\rangle\big| = \big|\langle x, \tfrac{Ar}{\|r\|}\rangle - \langle r, \tfrac{Ar}{\|r\|}\rangle\big| = \big|\langle x, \tfrac{Ar}{\|r\|}\rangle\big| \leq \frac{\|x\|\|Ar\|}{\|r\|} \leq \beta$$

by an application of the Cauchy–Schwarz inequality, and so we have

$$\|x - r\|^2 \leq \big|\langle x - r, \tfrac{r}{\|r\|}\rangle\big|^2 + \beta^2.$$

Noting that the radius β is quite small relative to the distance $\|r\|$, this allows the approximation $\|x - r\| \approx |\langle x - r, z\rangle|$, where $z := -r/\|r\|$ is the *bearing vector* between the aircraft and the scene; that is, z is the unit vector that determines the direction from the aircraft to the center of the scene. Essentially, this approximation makes use of the assumption $\|r\| \gg \beta$ to conclude that arcs of constant distance from the aircraft intersect the scene as straight lines perpendicular to its bearing vector z; such lines are referred to as lines of *constant range*, and so each distance in the direction of the bearing vector defines a unique line of constant range. In most applications of airborne SAR, this approximation is nearly sharp [48], and so we consider $\|x - r\| \approx |\langle x - r, z\rangle|$ for every $x \in D$.

Suppose the aircraft transmits a signal f, which is then reflected by the scene and returned to the aircraft as the signal g. Making the reasonable assumption that the speed of light far exceeds that of the aircraft, we know that f and g will travel the same path to and from any point $x \in D$; hence, both signals will travel the same distance, namely $\|x - r\|$. As a result, we expect the received signal to be given by

$$g(t) = \iint\limits_D f\big(t - \tfrac{2}{c}\|x - r\|\big)\rho(x)dx_1dx_2 \approx \iint\limits_D f\big(t - \tfrac{2}{c}\langle x - r, z\rangle\big)\rho(x)dx_1dx_2.$$

More generally, if the signal f is transmitted at a position r_1, then the received signal at any point r_2 is

$$g(t) \approx \iint\limits_D f\Big(t - \tfrac{1}{c}\big(\langle x - r_1, z_1\rangle + \langle x - r_2, z_2\rangle\big)\Big)\rho(x)dx_1dx_2, \tag{39}$$

where $z_1 := -r_1/\|r_1\|$, $z_2 := -r_2/\|r_2\|$, and it is assumed that $\|r_1\| \gg \beta$ and $\|r_2\| \gg \beta$, i.e., the transmitter and receiver are both sufficiently far from the scene.

For a fixed transmitter and receiver, it is unclear how to distinguish two points $x, y \in D$, $x \neq y$, given the signal g recorded over a period of time. The problem, which we discuss in the monostatic setting for simplicity, is two-fold: First, points along any line of constant range are necessarily indiscernible, since their radar signatures return to the source at exactly the same time. Hence, each recorded signal only contains information about the integrated reflectivity function along each line of constant range. The fix for this is the so-called *synthetic aperture*, which is the effective distance the aircraft travels while repeatedly imaging the scene. As the aircraft moves across the synthetic aperture, the lines of constant range rotate, and so each recorded signal encodes the integrated reflectivity function along a different set of lines through the scene [48]. Hence, wider synthetic apertures yield more information about lines of constant range.

The second problem is distinguishing points of differing distances from the aircraft. Although the radar signatures of such points return to the source at different times, the relative size of the scene (and the magnitude of the speed of light) cause them to overlap if the transmission duration is too small. Indeed, if the signal is of constant frequency, then it is impossible to tell two such points apart, regardless of the duration of the transmission. On the other hand, if the transmitted signal is a superposition of a range of frequencies (i.e., a *burst*), then it is possible to distinguish positions within the scene according to time of arrival, provided the duration of transmission is large enough. However, transmitting a burst is expensive and impractical due to power limitations.

To avoid this issue, a commonly used technique in airborne SAR is taking the transmitted signal f to be a *linear chirp* [48]. That is, let $f(\tau) := e^{2\pi\mathrm{i}(\frac{1}{2}v\tau^2+w\tau)}$, with instantaneous frequency

$$\frac{f'(\tau)}{2\pi\mathrm{i}f(\tau)} = \frac{2\pi\mathrm{i}(v\tau+w)e^{2\pi\mathrm{i}(\frac{1}{2}v\tau^2+w\tau)}}{2\pi\mathrm{i}e^{2\pi\mathrm{i}(\frac{1}{2}v\tau^2+w\tau)}} = v\tau + w.$$

Here, v is known as the *chirp rate* and w the *base frequency.* Notice that the instantaneous frequency of a chirp is linear in time; this enables points at different distances from the aircraft to be distinguished by extending the transmission duration (essentially amplifying relative distances) with a low power requirement. If the synthetic aperture is wide enough and contains enough transmission points, the resulting collection of line integrals of the reflectivity function is sufficient to distinguish all points within the scene [48].

Under the appropriate assumptions on the chirp rate of the transmitted signal, the reflected signal has a convenient form in terms of the two-dimensional Fourier transform $F\colon L^1(\mathbb{R}^2) \to L^\infty(\mathbb{R}^2)$ defined by

$$(Fh)(p,q) = \iint\limits_{\mathbb{R}^2} h(x,y)e^{-2\pi\mathrm{i}\langle(p,q),(x,y)\rangle}dxdy$$

for every $h \in L^1(\mathbb{R}^2)$. One way in which this result is formulated is by viewing each line integral of the reflectivity function as a one-dimensional Fourier transform and applying the Projection-Slice Theorem (see Section 2.3.2 in [48]). Alternatively, one could work entirely in two-dimensions using linear operators. The general result, proven by the latter approach, is given in the following:

Fact 5.1. *Let $r_1, r_2 \in \mathbb{R}^2$, with $\|r_1\| \gg \beta$ and $\|r_2\| \gg \beta$, and f be a chirp with chirp rate v and base frequency w. Furthermore, suppose that the signal f is emitted at position r_1 and the reflected signal g is received at r_2, as in* (39)*. If $v \ll \frac{c^2}{\beta(\|r_1\|+\|r_2\|)^2}$, then*

$$g(t) \approx f(t) e^{-2\pi \mathrm{i}(vt+w)\frac{1}{c}(\|r_1\|+\|r_2\|)} (F\rho)\left(\tfrac{1}{c}(vt+w)(z_1+z_2)\right).$$

Before proving this, we require some definitions. Consider the modulation and translation operators E and T defined respectively by $(Eh)(t) = e^{2\pi \mathrm{i} t} h(t)$ and $(Th)(t) = h(t-1)$ for any signal $h\colon \mathbb{R} \to \mathbb{C}$. These two operators act on a chirp $f(\tau) = e^{2\pi \mathrm{i}(\frac{1}{2}v\tau^2 + w\tau)}$ in a particular way: For any $a \in \mathbb{R}$,

$$\begin{aligned}
(T^a f)(\tau) = f(\tau - a) &= e^{2\pi \mathrm{i}\left(\frac{1}{2}v(\tau-a)^2 + w(\tau - a)\right)} \\
&= e^{2\pi \mathrm{i}\left((\frac{1}{2}va^2 + wa) - av\tau + (\frac{1}{2}v\tau^2 + w\tau)\right)} \\
&= e^{2\pi \mathrm{i}(\frac{1}{2}va^2 - wa)} e^{2\pi i(-avt)} f(\tau) = e^{2\pi \mathrm{i}(\frac{1}{2}va^2 - wa)} (E^{-av} f)(\tau).
\end{aligned}$$

Proof of Fact 5.1. Recall the form of the received signal (39). Defining the parameter $t_x := \frac{1}{c}(\langle x - r_1, z_1\rangle + \langle x - r_2, z_2 \rangle)$, we obtain

$$\begin{aligned}
g(t) &\approx \iint\limits_D f(t - t_x)\rho(x)dx_1dx_2 = \iint\limits_D \left(T^{t_x} f\right)(t)\rho(x)dx_1dx_2 \\
&= \iint\limits_D e^{2\pi \mathrm{i}\left(\frac{1}{2}vt_x^2 - wt_x\right)} \left(E^{-vt_x} f\right)(t)\rho(x)dx_1dx_2 \\
&= \iint\limits_D e^{\pi \mathrm{i} v t_x^2} e^{-2\pi \mathrm{i} w t_x} e^{-2\pi \mathrm{i}(vt)t_x} f(t)\rho(x)dx_1dx_2.
\end{aligned}$$

Now consider the signal

$$\hat{g}(t) := \iint\limits_D e^{-2\pi\mathrm{i}wt_x} e^{-2\pi\mathrm{i}(vt)t_x} f(t)\rho(x)dx_1dx_2.$$

Then

$$\begin{aligned}|\hat{g}(t) - g(t)| &= \left|\iint\limits_D \big(1 - e^{\pi\mathrm{i}vt_x^2}\big) e^{-2\pi\mathrm{i}wt_x} e^{-2\pi\mathrm{i}(vt)t_x} f(t)\rho(x)dx_1dx_2\right| \\ &\leq \iint\limits_D \big|1 - e^{\pi\mathrm{i}vt_x^2}\big||\rho(x)|dx_1dx_2 \\ &\leq \left(\iint\limits_D \big|1 - e^{\pi\mathrm{i}vt_x^2}\big|^2 dx_1dx_2\right)^{1/2} \left(\iint\limits_D |\rho(x)|^2 dx_1dx_2\right)^{1/2},\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz. By considering a Taylor series, it can be seen that $|1 - e^{\mathrm{i}\theta}|^2 = 2(1 - \cos(\theta)) \leq \theta^2$ for any $\theta \in \mathbb{R}$, and so it follows that

$$|\hat{g}(t) - g(t)| \leq \|\rho\|_{L^2} \left(\iint\limits_D \big(\pi vt_x^2\big)^2 dx_1dx_2\right)^{1/2}.$$

To bound the parameter t_x, first note that the Cauchy–Schwarz inequality yields

$$\begin{aligned}t_x &= \tfrac{1}{c}\big(\langle x - r_1, z_1\rangle + \langle x - r_2, z_2\rangle\big) \\ &= \tfrac{1}{c}\left(\left\langle x - r_1, \tfrac{-r_1}{\|r_1\|}\right\rangle + \left\langle x - r_2, \tfrac{-r_2}{\|r_2\|}\right\rangle\right) \\ &= \tfrac{1}{c}\left(\|r_1\| + \|r_2\| - \left\langle x, \tfrac{r_1}{\|r_1\|}\right\rangle - \left\langle x, \tfrac{r_2}{\|r_2\|}\right\rangle\right) \\ &\leq \tfrac{1}{c}\left(\|r_1\| + \|r_2\| + 2\|x\|\right),\end{aligned}$$

and so, over the region D, we obtain $0 \leq t_x \leq \frac{2}{c}(\|r_1\| + \|r_2\|)$. Thus,

$$|\hat{g}(t) - g(t)| \leq \|\rho\|_{L^2}\big(\tfrac{4\pi v}{c^2}(\|r_1\| + \|r_2\|)^2\big)(\pi\beta^2)^{1/2} = \tfrac{4\beta v}{c^2} \cdot \pi^{3/2}\|\rho\|_{L^2}(\|r_1\| + \|r_2\|)^2.$$

Recalling the assumption $v \ll \frac{c^2}{\beta(\|r_1\|+\|r_2\|)^2}$, this bound implies that the signal $\hat{g}$ approximates the received signal, and so we may write

$$g(t) \approx \hat{g}(t) = f(t) \iint\limits_D e^{-2\pi \mathrm{i}(vt+w)t_x} \rho(x) dx_1 dx_2.$$

Since $t_x = \frac{1}{c}(\|r_1\| + \|r_2\| + \langle x, z_1 + z_2 \rangle)$, it follows that

$$\begin{aligned} g(t) &\approx f(t) e^{-2\pi \mathrm{i}(vt+w)\frac{1}{c}(\|r_1\|+\|r_2\|)} \iint\limits_{\mathbb{R}^2} e^{-2\pi \mathrm{i}(vt+w)\langle x, z_1+z_2 \rangle} \rho(x) dx_1 dx_2 \\ &= f(t) e^{-2\pi \mathrm{i}(vt+w)\frac{1}{c}(\|r_1\|+\|r_2\|)} (F\rho) \left(\tfrac{1}{c}(vt+w)(z_1+z_2) \right), \end{aligned}$$

completing the proof. □

The implication of Fact 5.1 is that the received signal is a modulated version of the original transmitted signal times the Fourier transform of the reflectivity function along the line through the scene which bisects the lines defined by the bearing vectors z_1, z_2. For a single aircraft with position r, the lemma yields

$$g(t) = f(t) e^{-2\pi \mathrm{i}(vt+w)\frac{2}{c}\|r\|} (F\rho) \left(\tfrac{2}{c}(vt+w)z \right), \tag{40}$$

provided the chirp rate of f is chosen such that $v \ll \frac{c^2}{2\beta\|r\|}$. Thus, as the aircraft moves across the synthetic aperture, each received signal encodes the Fourier transform of the reflectivity function along a different line through the scene, namely, those lines defined by the bearing vectors of the transmission locations. Consequently, the image of the scene can be reconstructed if enough samples are taken to enable the entire Fourier transform to be built via interpolation. This is how SAR is implemented in practice.

The issue that arises with this implementation is the presence of phase error. This phase error is actually a consequence of relative uncertainty in the distance $\|r\|$ at each transmission location, which directly impacts the modulation and phase factors which precede the Fourier transform in (40). Small fluctuations in this distance, which are relatively common due to factors such as aircraft performance, weather, wind, and pilot skill, result in noticeable phase differences between the received signals [18]. As a result, each line of the Fourier transform of the reflectivity function obtained is skewed by an independent modulation and phase factor. Since estimating the modulation is possible by conventional methods [48], one could use interpolation to obtain the complete Fourier transform of the reflectivity function (up to a global phase factor) if these phase factors were all the same, which would then enable image reconstruction. Unfortunately, the uncertainties in target distance are typically uncorrelated, and so any image reconstruction algorithm must first clear this hurdle.

In practice, modern inertial navigation systems are capable of keeping the uncertainty in the distance $\|r\|$ small enough to render the modulating term in (40) negligible [48]. The remaining phase error, however, cannot be eliminated in this way. Two methods are used to deal with the phase error: One method is to use motion sensors on the aircraft to detect fluctuations in the flight path, which can be used to compute a correction to $\|r\|$ and determine the phase errors directly [63]. The second method, known as *autofocus*, is more widely used—it estimates the phase errors from the raw data under a priori assumptions on the image model [44]. Common autofocus algorithms include phase gradient and map drift [21, 28]; of these, phase gradient is usually preferred since the phase estimation portion of the algorithm is known to be optimal in a maximum-likelihood sense [48].

As an alternative, if we introduce additional information to the data set, it may be possible to completely determine the phase errors during image reconstruction. Suppose we insert a second aircraft into the SAR imaging process. Let the two aircraft have positions r_1 and r_3, with bearing vectors z_1 and z_3, respectively (position r_2 and bearing vector z_2 will be introduced later). Suppose aircraft i transmits the chirp $f(\tau) = e^{2\pi \mathrm{i}(\frac{1}{2}v\tau^2 + w\tau)}$ with chirp rate satisfying $v \ll \frac{c^2}{\beta(\|r_i\|^2 + \|r_j\|^2)}$. Then, by

Fact 5.1, we expect the received signal at aircraft j to be of the form

$$g_{i\to j}(t) := f(t)e^{-2\pi\mathrm{i}(vt+w)\frac{1}{c}(\|r_i\|+\|r_j\|)}(F\rho)\left(\tfrac{1}{c}(vt+w)(z_i+z_j)\right); \tag{41}$$

when $i \neq j$, this type of signature is characteristic of *bistatic* radar.

Consider the three received signals $g_{1\to 1}$, $g_{3\to 3}$, and $g_{1\to 3}$ obtained from (41). If we combine these signals in a particular way, we can eliminate the modulations and phase factors:

$$\begin{aligned}
&g_{1\to 1}(t)\overline{g_{1\to 3}(t)}^{\,2} g_{3\to 3}(t)\\
&\quad= f(t)e^{-2\pi\mathrm{i}(vt+w)\frac{2}{c}\|r_1\|}(F\rho)\left(\tfrac{2}{c}(vt+w)z_1\right)\\
&\qquad\times\left(\overline{f(t)e^{-2\pi\mathrm{i}(vt+w)\frac{1}{c}(\|r_1\|+\|r_3\|)}(F\rho)\left(\tfrac{1}{c}(vt+w)(z_1+z_3)\right)}\right)^2\\
&\qquad\times f(t)e^{-2\pi\mathrm{i}(vt+w)\frac{2}{c}\|r_3\|}(F\rho)\left(\tfrac{2}{c}(vt+w)z_3\right)\\
&\quad= |f(t)|^4(F\rho)\left(\tfrac{2}{c}(vt+w)z_1\right)\left(\overline{(F\rho)\left(\tfrac{1}{c}(vt+w)(z_1+z_3)\right)}\right)^2(F\rho)\left(\tfrac{2}{c}(vt+w)z_3\right).
\end{aligned}$$

The modulation and phase factors completely cancel, and what remains is simply the magnitude of the original signal times a product of Fourier transforms of the reflectivity function along lines through the scene defined by the unit vectors z_1, z_3 and their bisector $z_2 := \frac{z_1+z_3}{\|z_1+z_3\|}$. This interferometric effect bears some resemblance to that leveraged by Alexeev, Bandeira, Fickus and Mixon in [3] to solve the phase retrieval problem.

Before exploring this connection, it is important to note that the Fourier transforms of the reflectivity function in the combination above are taken along lines through the scene that are scaled differently (since $z_1 + z_3$ is not twice a unit vector). Hence, the parameterization of this line differs from those of the other lines, making them incompatible in Cartesian coordinates. Theoretically, this is not an issue; indeed, we may still obtain the Fourier transform of the reflectivity function via interpolation in polar coordinates. However, there is no fast algorithm for taking the inverse Fourier transform of the result. In order to take advantage of the fast Fourier transform (FFT), one must first convert the data to rectangular form, and so this scaling issue needs to be addressed. Monostatic SAR also encounters this problem, which can be solved by a process called *polar-to-rectangular resampling* [48]. Essentially, this process enables the data along any line of the Fourier transform of the reflectivity function to be properly scaled without influencing the integrity of the data. Hence, knowing $F\rho$ along a line through the scene with any scaling factor is enough to determine $F\rho$ along that same line with any other scaling factor. Denoting the ith slice of the Fourier transform $F\rho$ by

$$h_i(t) := (F\rho)\left(\tfrac{2}{c}(vt+w)z_i\right), \tag{42}$$

this allows the above expression to be rearranged:

$$h_1(t)\overline{h_2(t)}^{\,2}h_3(t) = \left(\overline{K(t)}\right)^{-2} g_{1\to1}(t)\overline{g_{1\to3}(t)}^{\,2}g_{3\to3}(t), \tag{43}$$

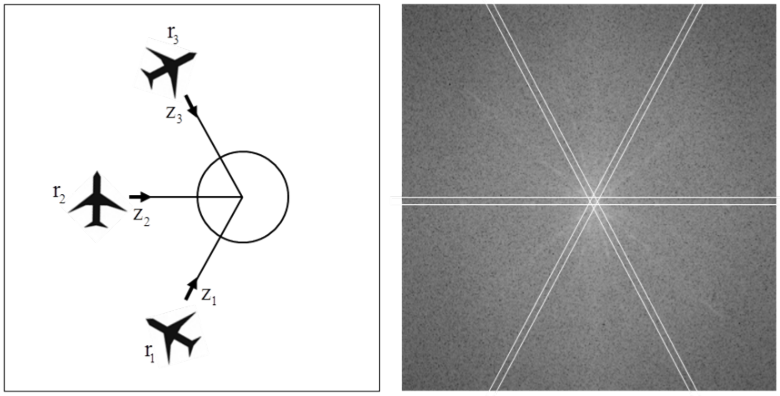
where $K(t)$ denotes the quantity

$$K(t) := |f(t)|^2 \frac{(F\rho)\left(\frac{1}{c}(vt+w)(z_1+z_3)\right)}{(F\rho)\left(\frac{1}{c}(vt+w)\left(\frac{z_1+z_3}{\|z_1+z_3\|}\right)\right)} = |f(t)|^2 \frac{(F\rho)\left(\frac{1}{c}(vt+w)(z_1+z_3)\right)}{h_2(t)}.$$

Note that $K(t)$ can be calculated by first estimating $(F\rho)(\frac{1}{c}(vt+w)(z_1+z_3))$ up to a global phase factor (by removing the modulation from $g_{1\to3}$ using conventional techniques [48]) and then using this to estimate $h_2(t)$ up to the same phase factor by dilation and translation. Hence, there is no ambiguity in the phase of $K(t)$.

At this point, it is useful to identify that each slice h_i of $F\rho$ can be obtained from the corresponding monostatic signal

$$g_{i\to i}(t) = f(t)e^{-2\pi\mathrm{i}(vt+w)\frac{2}{c}\|r_i\|}h_i(t).$$



(a) (b)

Figure 4: (a) A triple of aircraft positions (r_1, r_2, r_3) in a multistatic SAR scheme. (b) Using conventional techniques, the monostatic signals $g_{i\to i}$ associated with each bearing vector z_i produce the depicted slices h_i of the Fourier transform of the reflectivity function of the scene, up to distinct global phase factors ω_i. Meanwhile, the bistatic signal $g_{1\to 3}$ transmitted at r_1 and received at r_3 produces a dilated and translated version of the slice h_2. By Fact 5.2, this additional signal combines with the monostatic signals to determine the product $\omega_1\omega_2^{-2}\omega_3$ of unknown phases.

Again by conventional techniques [48], we can estimate $h_i(t)$ from $g_{i\to i}(t)$ up to a global phase factor ω_i. Thus, we get the estimates $\hat{h}_i(t) := \omega_i h_i(t)$ from the monostatic signal at position r_i. Expressing the received signals $g_{1\to 1}$, $g_{3\to 3}$, and $g_{1\to 3}$ in terms of these estimates, (43) implies

$$\omega_1\omega_2^{-2}\omega_3 = \frac{\hat{h}_1(t)\overline{\hat{h}_2(t)}^{\,2}\hat{h}_3(t)}{h_1(t)\overline{h_2(t)}^{\,2}h_3(t)}, \tag{44}$$

and so the bistatic signal $g_{1\to 3}$ (when combined with $g_{1\to 1}$, $g_{2\to 2}$, and $g_{3\to 3}$) enables one to recover a product of the phase factors of h_1, h_2, and h_3 (see Figure 4 for an illustration). We summarize this situation in the following fact:

Fact 5.2. *Pick $r_1, r_2, r_3 \in \mathbb{R}^2$ such that $r_1 + r_3$ is a positive scalar multiple of r_2, and let f be a chirp with chirp rate v and base frequency w satisfying the hypotheses of Fact 5.1. Suppose we obtain the signals $g_{1\to 1}$, $g_{2\to 2}$, $g_{3\to 3}$ as defined in* (41)*. Then conventional techniques [48] yield the estimates $\hat{h}_i(t) = \omega_i h_i(t)$ for each $i = 1, 2, 3$. Here, $h_i(t)$ denotes the ith slice of the desired Fourier transform* (42)*, and ω_i is an unknown phase factor. Furthermore, if we obtain $g_{1\to 3}$, then combining with the* *other signals according to* (43) *and* (44) *determines the product* $\omega_1\omega_2^{-2}\omega_3$ *of unknown phase factors.*

Since the magnitudes of the Fourier transform slices h_i can be obtained from the corresponding monostatic signals, determining these slices up to a single global phase requires determining the phase errors ω_i from products of the form (44). Notice that these products can be expressed in terms of relative phases: $\omega_1\omega_2^{-2}\omega_3 = (\omega_1\omega_2^{-1})(\omega_2\omega_3^{-1})^{-1}$. Defining the relative phases $\sigma_{1,2} := \omega_1\omega_2^{-1}$ and $\sigma_{2,3} := \omega_2\omega_3^{-1}$, we see that this quantity is itself a relative phase of relative phases: $\omega_1\omega_2^{-2}\omega_3 = \sigma_{1,2}\sigma_{2,3}^{-1}$. If this bistatic process is implemented while the two aircraft move across the synthetic aperture, it is possible to record such a quotient of relative phases for each triple (r_i, r_j, r_k) of locations realized by their flight paths. Extracting the individual phase errors from this collection of products is then possible via an algorithm used by Bandeira et al. in [3] to solve the phase retrieval problem, namely, *angular synchronization.*

5.2 Angular synchronization

The angular synchronization algorithm, as first introduced by Singer in [84], estimates a set of unknown phases using (noisy) measurements of relative phase. The idea is to organize the phase and relative phase information using a graph G, from which certain spectral methods enable the desired estimation with little computational burden. Furthermore, the algorithm is provably stable, provided the graph G is "nice" enough. Here, the proper notion of "nice" is in terms of the connectivity of G.

To set the stage, suppose we want to estimate a vector of phases $\omega := \{\omega_i\}_{i\in V} \subseteq \mathbb{C}$ for some finite set V given a set of relative phase measurements $\{\omega_i\omega_j^{-1}\}_{(i,j)\in E}$, where $E \subseteq V \times V$. Consider the simple graph $G = (V, E)$ so that each vertex in V represents an unknown phase and each edge in G a relative phase measurement. In particular, $i \in V$ represents the phase ω_i, while $(i, j) \in E$ if and only if $i, j \in V$ and we have the measurement $\omega_i \omega_j^{-1}$.

Notice that in the noiseless case, it is possible to recover the vector ω (up to a global phase) if and only if G is connected [84]. Indeed, in this case one may choose a spanning tree T of G and compute each phase ω_i by propagating from vertex to vertex using edges in T. To be clear, by propagate we mean multiply the phase at vertex i by the relative phase $\omega_j \omega_i^{-1}$ to obtain the phase at vertex j. The global phase ambiguity arises from the choice of starting phase, i.e., the phase assigned to the root of T [84]. Unfortunately, this approach does not work in the noisy case; in fact, this method actually compounds the noise by adding another noise term at each vertex.

Instead of propagating the relative phases along only the edges in T, suppose that we propagate along every edge in E. If there are no cycles in G, then this exhibits the same problem as above. However, cycles provide a means of "noise cancellation." To see this, suppose C is a cycle in G. Choosing two vertices $i, j \in C$, there are two paths in C from i to j. Thus, we can choose to propagate the phase at vertex i along either of these paths to obtain the phase at vertex j, knowing that we expect to obtain the same result regardless of which path we choose. Due to noise, it is unlikely that one will obtain the same result for both paths, but the two differing phases at vertex j provide a means for comparison. In particular, if we take $j = i$, then the sum of the phase errors along C must be zero modulo 2π; this cancellation provides a means of combating the noise. Hence, the more cycles there are in G, the more such comparisons one can make. This observation suggests that graphs with many cycles are desirable, a quality which is present in highly connected graphs. This idea is what led Bandeira et al. to exploit an (optimally connected) expander graph to solve the phase retrieval problem using $O(M \log M)$ intensity measurements [3].

What follows is a more in-depth discussion of angular synchronization, but first we require some definitions from spectral graph theory. For a simple, connected graph with n vertices, let A denote the adjacency matrix and D the diagonal matrix of vertex degrees, $\{d_i\}_{i=1}^n \subseteq \mathbb{N}$. The *Laplacian* of the graph is defined to be the matrix $L := I - D^{-1/2}AD^{-1/2}$, which is positive semidefinite since for any $x \in \mathbb{C}^n$ we have

$$x^*Lx = \|x\|^2 - \sum_{i=1}^n \sum_{j=1}^n \frac{x_i^* A[i,j] x_j}{\sqrt{d_i d_j}} \geq \|x\|^2 - \sum_{i=1}^n \sum_{j=1}^n x_i^* x_j = 0$$

with equality when x is the vector whose entries are all 1; here, the inequality follows from the fact that $d_i d_j \geq 1$ whenever $A[i,j] = 1$. Hence, the spectrum of L satisfies $0 = \lambda_1 \leq \cdots \leq \lambda_n$. The second eigenvalue λ_2 of the Laplacian is known as the *spectral gap* of the graph. As discussed in [3], a highly connected graph necessarily has a large spectral gap.

Returning to the graph $G = (V, E)$ defined above, let $\varepsilon := \{\varepsilon_{ij}\}_{(i,j)\in E}$ denote a vector of adversarial noise terms. Since we seek to propagate phase along edges, note that the noisy relative phase $\omega_i \omega_j^{-1} + \varepsilon_{ij}$ is associated with a direction, namely, propagation from vertex i to vertex j. For the reverse direction, we take the noisy relative phase $\omega_j \omega_i^{-1} + \varepsilon_{ji} = \overline{\omega_i \omega_j^{-1} + \varepsilon_{ij}}$. Let A_1 denote the weighted adjacency matrix of G, defined entrywise by

$$A_1[i,j] := \frac{\omega_i \omega_j^{-1} + \varepsilon_{ij}}{|\omega_i \omega_j^{-1} + \varepsilon_{ij}|}. \tag{45}$$

Note that A_1 is self-adjoint and each entry $A_1[i,j]$ is an approximation of the relative phase $\omega_i \omega_j^{-1}$. Considering this, it makes sense that we may obtain the vertex phases $\{\omega_i\}_{i\in V}$ from A_1 by solving the minimization problem

$$\hat{\omega} = \operatorname*{arg\,min}_{\{\omega_i\}_{i\in V} \subseteq \mathbb{T}} \sum_{(i,j)\in E} |\omega_i - A_1[i,j]\omega_j|^2. \tag{46}$$

One method of (approximately) solving this problem is *angular synchronization*, which is summarized in Algorithm 3 (cf. [3]).

Algorithm 3 Angular synchronization

```
Input: Graph $G = (V, E)$, noisy relative phases $\omega_i\omega_j^{-1} + \varepsilon_{ij}$ for every $(i, j) \in E$
Output: Vector of phases $\{\omega_i\}_{i\in V}$
  Let $A_1$ denote the weighted adjacency matrix of $G$, defined entrywise in (45)
  Let $D$ denote the diagonal matrix of vertex degrees $\{d_i\}_{i\in V}$
  Compute the matrix $L_1 \leftarrow I - D^{-1/2}A_1D^{-1/2}$
  Compute the eigenvector $u$ corresponding to the smallest eigenvalue of $L_1$
  Ouput $\omega_i = u_i/|u_i|$ for every $i \in V$
```

The matrix $L_1 := I - D^{-1/2}A_1D^{-1/2}$ in Algorithm 3 is known as the *connection Laplacian* of G, and bears resemblance to the Laplacian L. Note that Algorithm 3 suggests that the solution $\hat{\omega}$ to (46) is (approximately) the vector whose entries are normalized versions of the entries of the eigenvector corresponding to the smallest eigenvalue. To see this, we will need the help of the following elementary result from graph theory:

Lemma 5.3 (Degree-Sum Formula)**.** *Let $G = (V, E)$ be a simple graph and denote by d_i the degree of vertex $i \in V$. Then $\sum_{i\in V} d_i = 2|E|$.*

Proof. Consider the subset of $S \subseteq V \times E$ consisting of all pairs of vertices and their incident edges, namely $S := \{(i, (j, k)) : i, j, k \in V,\ j = i \text{ or } k = i\}$. Since the degree of a vertex counts its incident edges, every $i \in V$ contributes d_i elements to S. On the other hand, every edge $(i, j) \in E$ determines exactly two elements of S, namely $(i, (i, j))$ and $(j, (i, j))$. Hence, we have $\sum_{i\in V} d_i = |S| = 2|E|$, as desired. □

We now provide some intuition behind the use of eigenvectors in Algorithm 3. Expanding the sum in (46), we see that

$$\begin{aligned}\sum_{(i,j)\in E} |\omega_i - A_1[i, j]\omega_j|^2 &= \sum_{(i,j)\in E} \Big(|\omega_i|^2 - 2\,\mathrm{Re}(\omega_i^{-1}A_1[i, j]\omega_j) + |A_1[i, j]\omega_j|^2\Big)\\ &= 2\sum_{(i,j)\in E} \Big(1 - \mathrm{Re}(\omega_i^{-1}A_1[i, j]\omega_j)\Big).\end{aligned}$$

By Lemma 5.3, the first part of this expression becomes

$$2|E| = \sum_{i\in V} d_i = \sum_{i\in V} \omega_i^{-1} d_i \omega_i = \omega^* D\omega. \tag{47}$$

Meanwhile, the second term yields

$$\mathrm{Re}\Big(2 \sum_{(i,j)\in E} \omega_i^{-1} A_1[i,j]\omega_j\Big) = \sum_{i\in V}\sum_{j\in V} \omega_i^{-1} A_1[i,j]\omega_j = \omega^* A_1 \omega,$$

and so (46) may be written $\hat{\omega} = \arg\min_{\omega\in\mathbb{T}^{|V|}} \omega^*(D - A_1)\omega$. Noting from (47) that $\omega^* D\omega$ does not vary with ω, this is equivalent to solving

$$\arg\min_{\omega\in\mathbb{T}^{|V|}} \frac{\omega^*(D-A_1)\omega}{\omega^* D\omega},$$

which, since $D - A_1$ is self-adjoint, is a Rayleigh quotient in terms of the connection Laplacian of G:

$$\frac{\omega^*(D-A_1)\omega}{\omega^* D\omega} = \frac{(D^{1/2}\omega)^*(I - D^{-1/2}A_1 D^{-1/2})(D^{1/2}\omega)}{\|D^{1/2}\omega\|^2} = \frac{(D^{1/2}\omega)^* L_1 (D^{1/2}\omega)}{\|D^{1/2}\omega\|^2}.$$

The minimum value of this Rayleigh quotient over all $\omega \in \mathbb{C}^{|V|}$ is the smallest eigenvalue of L_1, attained when $u := D^{1/2}\omega$ is the corresponding eigenvector. Since the ith coordinate of $D^{1/2}\omega$ is $\sqrt{d_i}\omega_i$, it follows that $\omega_i = u_i/\sqrt{d_i}$, is the optimal choice for the relaxed form of (46). This does not necessarily have unit-modulus entries, however, and so we normalize: $\hat{\omega}_i := u_i/|u_i|$.

To be clear, since Algorithm 3 produces an estimate $\hat{\omega}$ for the vector of unknown phases using only relative phases, any estimate $e^{i\theta}\hat{\omega}$, $\theta \in \mathbb{R}$, is also a viable estimate; this is reflected in the fact that the eigenvector u is unique up to a complex scalar. Thus, angular synchronization yields the desired phase vector up to a global phase ambiguity. Bandeira et al. [3] prove a stability guarantee for Algorithm 3 in terms of the spectral gap λ_2 of the graph G. As such, we don't lose much in the relaxation provided the underlying graph is sufficiently connected. We state their result here (without proof) for completeness:

Theorem 5.4 (Theorem 6.3 in [3]). *Consider a graph $G = (V, E)$ with spectral gap $\lambda_2 > 0$ and define $\|\theta\|_{\mathbb{T}} := \min_{k\in\mathbb{Z}} |\theta - 2\pi k|$ for all $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Furthermore, let A_1 denote the weighted adjacency matrix of G, defined entrywise in* (45)*. Then Algorithm 3 outputs the estimate $\hat{\omega} \in \mathbb{C}^{|V|}$ with unit-modulus entries such that, for some $\theta \in \mathbb{R}/2\pi\mathbb{Z}$,*

$$\sum_{i\in V} \| \arg(\hat{\omega}_i) - \arg(\omega_i) - \theta\|_{\mathbb{T}}^2 \leq \frac{C\|\varepsilon\|^2}{P^2\lambda_2^2}$$

where $P := \min_{(i,j)\in E} |\omega_i\omega_j^{-1} + \varepsilon_{ij}|$ and C is a universal constant.

5.3 Formulating the phase error problem with graphs

Recall the concluding discussion of Section 5.1, in which we combined several received radar signals with their estimates to produce a product of their phase errors (44). In particular, this product of phase errors is a quotient of relative phases. With angular synchronization in hand, we will see that this nesting of relative phases suggests a way to use angular synchronization to recover the phase errors.

First, we show how to organize the available phase information using graphs. To this end, consider the graph $G = (V, E)$ for some finite set V. Let $\{\omega_i\}_{i\in V}$ be a vector of unknown phase errors and $\{\omega_i\omega_j^{-1}\}_{(i,j)\in E}$ a set of unknown relative phases. Now define a second graph $G' = (V', E')$ such that $V' = E$, and consider the set of known relative phases $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$, where $\sigma_{i,j} := \omega_i\omega_j^{-1}$ for every $(i,j) \in E$. Note from Fact 5.2 that we have $\omega_i\omega_j^{-2}\omega_k = \sigma_{i,j}\sigma_{j,k}^{-1}$ for every $((i,j),(j,k)) \in E'$, and so G and G' together encode a nested set of relative phases. Hence, one may seek to first apply Algorithm 3 on G' to determine the edge measurements for G (up to a global phase), at which point applying the algorithm a second time on G would determine the phase errors $\{\omega_i\}_{i\in V}$ (up to yet another global phase). Note that this

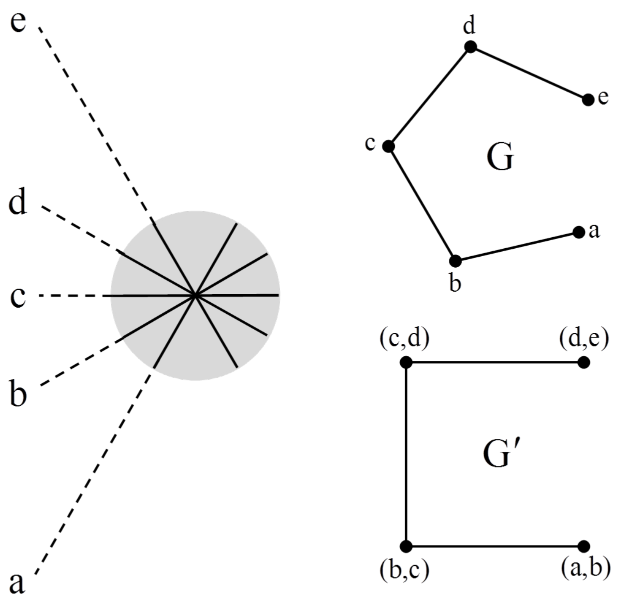


Figure 5: Two aircraft imaging a scene of interest (left) using bistatic SAR techniques at three different times. As in Example 5.5, at the first time instant the aircraft are located at positions a and c, while the second and third time instances correspond to location pairs (b, d) and (c, e), respectively. To form the corresponding graphs $G = (V, E)$ and $G' = (V', E')$ (right), note that Fact 5.2 states that, for each location pair, we receive relative phase information relating the positions to their bisector. Hence, each location pair contributes an edge to G'; for instance, the pair (a, c) dictates that $(a, b) \leftrightarrow (b, c)$. Furthermore, each of the vertices in G' corresponds to an edge in G. Here, two vertices are adjacent if one is the bisector of a location pair that the other belongs to. In this way, edges in G represent relative phases of the form $\omega_i\omega_j^{-1}$, while edges in G' encode the recorded quotients of relative phases $\sigma_{ij}\sigma_{j,k}^{-1}$. Angular synchronization provides a means of retrieving $\{\omega_i\omega_j^{-1}\}_{(i,j)\in V'}$ up to a global phase from $\{\sigma_{ij}\sigma_{j,k}^{-1}\}_{((ij),(jk))\in E'}$. Encoding the resulting vertex measurements on the edges in G then allows $\{\omega_i\}_{i\in V}$ to be obtained (up to a global phase and partial modulation) by a second application of angular synchronization.

process relies heavily on the connectivity of G', which is inherited from properties of G. As we will see, connectivity in G' is not easily obtained. Regardless, the idea of using angular synchronization on both graphs motivates the approach of this section.

Example 5.5. Suppose two aircraft image the scene in Figure 5 along the synthetic aperture from point a to point e. In particular, the two aircraft image the scene with location pairs (a, c), (b, d), and (c, e) (meaning the first aircraft is at position a when the second is at c, and so on). Based on Fact 5.2, we record a product of phase errors of the form $\sigma_{i,j}\sigma_{j,k}^{-1}$ at each location. To organize this information, we take the graphs $G = (V, E)$ and $G' = (V', E')$, depicted in Figure 5, such that

$$V = \{a, b, c, d, e\}, \qquad E = \big\{(a, b), (b, c), (c, d), (d, e)\big\}$$

and

$$\begin{aligned} V' &= E = \big\{(a,b),(b,c),(c,d),(d,e)\big\}, \\ E' &= \Big\{\big((a,b),(b,c)\big),\big((b,c),(c,d)\big),\big((c,d),(d,e)\big)\Big\}. \end{aligned}$$

In this way, each vertex in V represents an unknown phase error, each edge in E (equivalently, vertex in V') represents an unknown relative phase between phase errors, and each edge in E' a known quotient of such relative phases obtained using Fact 5.2. Thus, performing angular synchronization on G' and then on G will return an estimate of the phase errors with two degrees of ambiguity. □

Note that the graphs G and G' of Example 5.5 are both nearly cyclic; indeed, both cycles would be completed if the location pair (d,a) were added to the system. For this reason, suppose we allow two aircraft imaging a scene to follow a circular path enclosing the target, maintaining their relative positions throughout. Depending on the distance between the aircraft, the graphs generated by the bistatic phase errors are then cycles, where the number of vertices is equal to the number of location pairs at which the scene is imaged. Allowing for multiple pairs of aircraft to perform such a maneuver along this circular path then creates multiple cycles in G, each creating a separate cycle in G'. Hence, more planes effectively increases the connectivity of G, which means the second application of Algorithm 3 will be more stable by Theorem 5.4. It is for this reason that we focus on a particular class of graphs, namely, *circulant* graphs:

Definition 5.6. *A simple graph $G = (V, E)$ is said to be circulant if its adjacency matrix A is a circulant matrix. That is, there exists a mapping $\alpha\colon \mathbb{Z}_{|V|} \to \{0,1\}$ such that $A[i,j] = \alpha(j-i)$ for all $i, j \in \{0,\dots,|V|-1\}$.*

Circulant graphs enjoy a lot of useful mathematical properties. For example, circulant graphs whose vertex sets are of prime order have a convenient structure in terms of the cycles they contain.

Lemma 5.7. *Let G be a simple, circulant graph with n vertices. If n is an odd prime, then G decomposes into copies of the n-cycle.*

Proof. Let A denote the adjacency matrix of G. Then, because G is circulant, there exists a mapping $\alpha\colon \mathbb{Z}_n \to \{0,1\}$ such that $A[i,j] = \alpha(j-i)$ for all $i,j \in \{0,\dots,n-1\}$. Since the adjacency matrix of a simple graph is symmetric, we also have $\alpha(x) = \alpha(-x)$ for all $x \in \mathbb{Z}_n$. Moreover, the fact that the diagonal entries of A must vanish implies $\alpha(0) = 0$. Let $\delta_i\colon \mathbb{Z}_n \to \{0,1\}$ such that

$$\delta_i(x) := \begin{cases} 1 & \text{if } x = i \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in \{0,\dots,n-1\}$. Then α decomposes as

$$\alpha(x) = \sum_{i\in\mathbb{Z}_n} \alpha(i)\delta_i(x) = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} \alpha(i)(\delta_i + \delta_{-i})(x), \tag{48}$$

where the last equality follows from the assumption that n is odd. Now consider the set $S := \{1 \le s \le \lfloor \frac{n}{2} \rfloor : \alpha(s) = 1\}$ and note that, by the properties of α listed above, (48) may be written as

$$\alpha(x) = \sum_{s\in S} \alpha(s)(\delta_s + \delta_{-s})(x) = \sum_{s\in S} (\delta_s + \delta_{-s})(x).$$

Thus, defining the matrices $\{A_s\}_{s\in S}$ entrywise by $A_s[i,j] := (\delta_s + \delta_{-s})(j-i)$, we can decompose A as follows:

$$A[i,j] = \alpha(j-i) = \sum_{s\in S} (\delta_s + \delta_{-s})(j-i) = \sum_{s\in S} A_s[i,j].$$

We claim that each A_s is the adjacency matrix of an n-cycle, from which it follows that G decomposes into $|S|$ copies of C_n. To prove this claim, note that for any $s \in S$,

$A_s[i,j] = 1$ whenever $j - i = \pm s$. Hence, consider the vertex labeling $\phi\colon \mathbb{Z}_n \to \mathbb{Z}_n$ defined by $\phi(i) := s^{-1}i$ for all $i \in \mathbb{Z}_n$, where s^{-1} is the multiplicative inverse of s in $\mathbb{Z}_n$ (note that this is a well-defined bijection since s is nonzero and n is prime, implying $\mathbb{Z}_n$ is a field and s is invertible). Furthermore, since ϕ is linear, $j - i = \pm s$ if and only if

$$\phi(j) - \phi(i) = \phi(j - i) = \phi(\pm s) = \pm\phi(s) = \pm 1,$$

meaning that $A_s[i,j] = 1$ if and only if $\phi(j) - \phi(i) = \pm 1$. Thus, each A_s is isomorphic (by the corresponding ϕ) to the standard n-cycle, as desired. □

Now that we have a better understanding of circulant graphs, we apply this understanding to solve the phase error problem in the noiseless case. To this end, let $G = (V, E)$ be circulant with n vertices, where n is an odd prime. Also, consider the graph $G' = (V', E')$ where $V' = E$ and $(i,j), (k,\ell) \in V'$ are adjacent if and only if they share a neighbor in a common n-cycle in the decomposition of G given by Lemma 5.7. Thus, edges in E' are of the form $((i,j),(j,k))$ for $i,j,k \in C_n \subseteq G$. For instance, one way in which this can be implemented for multistatic SAR is to allow $c+1$ aircraft to circle a target scene, each performing monostatic SAR; if the first aircraft in the formation transmits a bistatic signal for each of the other aircraft to receive, then the resulting graph G will be circulant with c cycles.

To better develop our understanding of the phase error problem with circulant graphs, we first focus on the case where G is a cycle. To this end, suppose $\theta\colon V \to \mathbb{R}/\mathbb{Z}$ is any function on the vertices of an n-cycle in G and consider the function $\theta'\colon V' \to \mathbb{R}/\mathbb{Z}$ defined by $\theta'((i,j)) := \theta(j) - \theta(i)$ for every $(i,j) \in V'$. To be clear, θ' encodes differences (modulo 1) in the value of θ at adjacent vertices in G, and so resembles a finite difference. Similarly, consider the function $\theta''\colon E' \to \mathbb{R}/\mathbb{Z}$ defined by $\theta''(((i,j),(k,\ell))) := \theta'((k,\ell)) - \theta'((i,j))$, which encodes the same types of differences in the value of θ' at adjacent vertices in G'. Since we may identify V

with $\mathbb{Z}_n$ in such a way that $i, j \in V$ are adjacent whenever $j - i = \pm 1 \bmod n$, it is possible to redefine θ, θ', and θ'' as functions on $\mathbb{Z}_n$. In particular, we see that θ' and θ'' act on any $x \in \mathbb{Z}_n$ such that $\theta'(x) = \theta(x+1) - \theta(x)$ and $\theta''(x) = \theta'(x+1) - \theta'(x)$.

Note that the function θ may be made to represent a complex phase by considering the unit-modulus number $e^{2\pi \mathrm{i}\theta(x)}$ for any $x \in \mathbb{Z}_n$. With this in mind, θ' and θ'' analogously represent the type of nested relative phases we are interested in. (Indeed, we switch to θ-notation here so that we can appeal to intuitions of finite-differences and integration. Note that θ lies in $[0, 1)$ instead of $[0, 2\pi)$ due to the choice of normalization.) Thus, it is reasonable to ask whether it is possible to determine θ (up to some sort of global ambiguity having two degrees of freedom) given only the values of θ''. This leads to the following lemma:

Lemma 5.8. *For any function $\theta \colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$, let $\theta' \colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$ and $\theta'' \colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$ be defined by $\theta'(x) := \theta(x+1) - \theta(x)$ and $\theta''(x) := \theta'(x+1) - \theta'(x)$. Then the values $\{\theta''(x)\}_{x \in \mathbb{Z}_n}$ determine an estimate $\hat{\theta}$ such that $\hat{\theta}(x) = z + \frac{kx}{n} + \theta(x)$ for some $z \in \mathbb{R}/\mathbb{Z}$ and $k \in \{0, \ldots, n-1\}$.*

Considering the above discussion, the implication of this result is that knowing the relative phase measurements represented by the edges in G' determines the phases encoded at the vertices of any cycle in G up to a modulation and a global phase factor, not unlike the result of two iterations of angular synchronization alluded to in Example 5.5. Indeed, the processes $\theta'' \mapsto \hat{\theta}'$ and $\hat{\theta}' \mapsto \hat{\theta}$ are just angular synchronization in the noiseless case. To clarify, the estimate produced by Lemma 5.8 yields

$$e^{2\pi \mathrm{i}\hat{\theta}(x)} = e^{2\pi \mathrm{i} z} e^{2\pi \mathrm{i} kx/n} e^{2\pi \mathrm{i}\theta(x)}$$

for some $z \in \mathbb{R}/\mathbb{Z}$; hence, the values of $e^{2\pi \mathrm{i}\hat{f}(x)}$ are a modulation of the actual values of $e^{2\pi \mathrm{i} f(x)}$, multiplied by a global phase. The global phase is precisely the ambiguity associated with the second application of angular synchronization (on the graph G), while the modulation is a consequence of propagating the ambiguity from the first application of angular synchronization (on the graph G') through the second. In the noiseless case, this process is particularly clean; in fact, a single n-cycle in G is sufficient:

Corollary 5.9. *Let $G = (V, E)$ be an n-cycle and $G' = (V', E')$ where $V' = E$ and $(i, j), (k, \ell) \in V'$ are adjacent if and only if they share a vertex as edges in G. Furthermore, let $\{\omega_i\}_{i \in V}$ be a vector of unknown phase errors, $\{\omega_i\omega_j^{-1}\}_{(i,j)\in E}$ a set of unknown relative phases, and $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$ a set of known relative phases, where $\sigma_{i,j} := \omega_i\omega_j^{-1}$ for every $(i, j) \in E$. Then the measurements $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$ determine ω up to a modulation and a global phase factor.*

Proof. Let $g\colon V \to \mathbb{R}/\mathbb{Z}$ such that $\omega_i = e^{2\pi \mathrm{i} g(i)}$ for every $i \in V$ and define the functions $g'\colon V' \to \mathbb{R}/\mathbb{Z}$, $g''\colon E' \to \mathbb{R}/\mathbb{Z}$ by

$$g'((i,j)) := g(j) - g(i) \qquad \text{and} \qquad g''\big(((i,j),(k,\ell))\big) := g'((k,\ell)) - g'((i,j)).$$

Note that these definitions imply

$$\begin{aligned} \omega_i\omega_j^{-1} &= e^{2\pi \mathrm{i} g'((j,i))} && \text{for all } (i,j) \in E, \\ \sigma_{i,j}\sigma_{j,k}^{-1} &= e^{2\pi \mathrm{i} g''(((k,j),(j,i)))} && \text{for all } ((i,j),(j,k)) \in E', \end{aligned}$$

where we have implicitly assigned a direction to each edge in G and G'. In particular, the set of measurements $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$ completely determines the values $\{g''(((i,j),(j,k)))\}_{((i,j),(j,k))\in E'}$.

Since G is an n-cycle, there exists a vertex labeling $\phi\colon V \to \mathbb{Z}_n$ such that $i, j \in V$ are adjacent if and only if $\phi(j) - \phi(i) = \pm 1 \bmod n$. Consider the functions

$\theta\colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$, $\theta'\colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$, and $\theta''\colon \mathbb{Z}_n \to \mathbb{R}/\mathbb{Z}$ defined earlier, so that

$$\begin{aligned}\theta(x) :=& g(\phi^{-1}(x)),\\ \theta'(x) :=& g'(\phi^{-1}(x+1)) - g'(\phi^{-1}(x)),\\ \theta''(x) :=& g''(\phi^{-1}(x+1)) - g''(\phi^{-1}(x)).\end{aligned}$$

Since the values $\{\theta''(x)\}_{x\in\mathbb{Z}_n}$ are also completely determined by the set of measurements $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$, applying Lemma 5.8 yields the estimate $\hat{\theta}(x) = z + \frac{kx}{n} + \theta(x)$ for some $z \in \mathbb{R}/\mathbb{Z}$ and $k \in \{0, \ldots, n-1\}$. Thus, we have

$$\begin{aligned}\hat{\omega}_i :=& e^{2\pi \mathrm{i}\hat{g}(i)} = e^{2\pi \mathrm{i}\hat{\theta}(\phi(i))} = e^{2\pi \mathrm{i}\left(z+\frac{k\phi(i)}{n}+\theta(\phi(i))\right)}\\ =& e^{2\pi \mathrm{i} z}e^{2\pi \mathrm{i} k\phi(i)/n}e^{2\pi \mathrm{i} g(i)} = e^{2\pi \mathrm{i} z}e^{2\pi \mathrm{i} k\phi(i)/n}\omega_i\end{aligned}$$

completing the proof. □

This result is not entirely surprising from the perspective of angular synchronization, since we have already seen how the recovery of unknown phases from noiseless relative phase measurements is possible provided the resultant graph is connected. Still, Corollary 5.9 illustrates the power of Lemma 5.8 as a means of analyzing the types of circulant graphs we are interested in.

Proof of Lemma 5.8. Since $\theta'(x+1) = \theta''(x) + \theta'(x)$ for any $x \in \{0, \ldots, n-1\}$, fixing the estimate $\hat{\theta}'(0)$ iteratively determines the set $\{\hat{\theta}'(x)\}_{x=0}^{n-1}$. To show that this set is consistent, we also require $\hat{\theta}'(0) = \theta''(n-1) + \hat{\theta}'(n-1)$. For this, note that an inductive argument yields

$$\hat{\theta}'(n-1) = \theta''(n-2) + \hat{\theta}'(n-2) = \sum_{k=0}^{n-2}\theta''(k) + \hat{\theta}'(0),$$

and so it suffices to show that

$$\theta''(n-1) = -\sum_{k=0}^{n-2} \theta''(k) = -\sum_{k=0}^{n-2} (\hat{\theta}'(k+1) - \hat{\theta}'(k))$$
$$= \hat{\theta}'(0) - \hat{\theta}'(n-1) = \theta'(0) - \theta'(n-1),$$

which follows from the definition of θ''. Hence, there exists $m \in \mathbb{R}/\mathbb{Z}$ such that $\hat{\theta}'(x) = m + \theta'(x)$ for all $x \in \mathbb{Z}_n$.

Similarly, since $\theta(x+1) = \theta'(x) + \theta(x)$ for any $x \in \{0, \ldots, n-1\}$, fixing the estimate $\hat{\theta}(0)$ iteratively determines the set $\{\hat{\theta}(x)\}_{x=0}^{n-1}$ in terms of the estimates $\{\hat{\theta}'(x)\}_{x=0}^{n-1}$; in particular, a telescoping sum in the definition of $\hat{\theta}(x)$ gives

$$\hat{\theta}(x) = mx + \theta(x) - (\theta(0) - \hat{\theta}(0)).$$

Hence, there exists $z \in \mathbb{R}/\mathbb{Z}$ such that $\hat{\theta}(x) = z + mx + \theta(x)$ provided the consistency relation $\hat{\theta}(0) = \theta'(n-1) + \hat{\theta}(n-1)$ is satisfied. To this end, an inductive argument first yields

$$\hat{\theta}(n-1) = \hat{\theta}'(n-2) + \hat{\theta}(n-2) = \sum_{k=0}^{n-2} \hat{\theta}'(k) + \hat{\theta}(0),$$

and so it suffices to show that

$$\hat{\theta}'(n-1) = -\sum_{k=0}^{n-1} \hat{\theta}'(k) = -\sum_{k=0}^{n-1} (m + \theta'(k))$$
$$= -m(n-1) - \sum_{k=0}^{n-1} (\theta(k+1) - \theta(k)) = -m(n-1) + \theta(0) - \theta(n-1).$$

By the definition of $\hat{\theta}'$, this only holds if $mn \equiv 0 \bmod 1$; that is, there must exist some integer k such that $m = k/n$. Therefore, we conclude that $\hat{\theta}(x) = z + \frac{kx}{n} + \theta(x)$ for all $x \in \mathbb{Z}_n$, where $z \in \mathbb{R}/\mathbb{Z}$ and $k \in \{0, \ldots, n-1\}$. □

Notice that the recovery process outlined in the proof of Corollary 5.9 fails when the measurements $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(j,k))\in E'}$ are corrupted by noise, and so we must seek a more stable solution. For this reason, we return to the more general case where G is a circulant graph with n vertices such that n is an odd prime, and we will represent the phase information in a way that is more compatible with angular synchronization. Since θ is any function taking vertices in V to real numbers modulo 1, let $\omega := \{\omega_i\}_{i\in V} \subseteq \mathbb{T}$ be a complex vector of phase errors such that $\omega_i := e^{2\pi \mathrm{i}\theta(i)}$ for every $i \in V$. Note that this construction is well-defined, and so we will treat the vector ω as a function from V to $\mathbb{T}$. Using the proof of Corollary 5.9 as motivation, consider the decomposition of G into n-cycles given by Lemma 5.7, so that we may orient each cycle to give every edge a direction. Now, redefine the function $\theta'\colon V' \to \mathbb{R}/\mathbb{Z}$ in terms of these directions; in particular, take $\theta'((i,j)) := \theta(j) - \theta(i)$ whenever $i \to j$ in G. Then

$$(\omega\omega^*)[i,j] = e^{2\pi\mathrm{i}(\theta(i)-\theta(j))} = e^{2\pi\mathrm{i}\theta'((j,i))}.$$

To similarly redefine θ'', let $B \subseteq E'$ denote the set of all ordered pairs of consecutive edges in a common n-cycle of G from the decomposition above and take $\theta''\colon B \to \mathbb{R}/\mathbb{Z}$ such that

$$\theta''\big(((i,j),(j,k))\big) := \theta'((j,k)) - \theta'((i,j)) = \theta(k) - 2\theta(j) + \theta(i).$$

Hence, we have

$$(\omega\omega^*)[i,j]\big((\omega\omega^*)[j,k]\big)^{-1} = e^{2\pi\mathrm{i}(\theta(i)-2\theta(j)+\theta(k))} = e^{2\pi\mathrm{i}\theta''(((k,j),(j,i)))},$$

and so such products between entries of $\omega\omega^*$ for edges in B may be used to estimate the entries of ω by applying Lemma 5.8.

To see this, note that Lemma 5.8 provides a means of producing an estimator for θ based on the given values of θ'' on a cycle. Specifically, for each cycle C one may obtain the estimate $\hat{\theta}(i) = z(C) + \frac{k(C)i}{n} + \theta(i)$ for some cycle-specific $z(C) \in \mathbb{R}/\mathbb{Z}$ and $k(C) \in \{0, \ldots, n-1\}$. In particular, for each cycle C in the decomposition of G, the values of θ'' determine the estimator $\hat{\theta}'((i,j)) = m(C) + \theta'((i,j))$ for all $(i,j) \in C$, where $m(C) \in \mathbb{R}/\mathbb{Z}$. With this in mind, consider the matrices X_C and X_{-C}, defined entrywise by

$$X_C[i,j] := \begin{cases} e^{-2\pi \mathrm{i} \hat{\theta}'((i,j))} & \text{if } i \to j \text{ in } C \\ 0 & \text{otherwise,} \end{cases}$$

$$X_{-C}[i,j] := \begin{cases} e^{2\pi \mathrm{i} \hat{\theta}'((j,i))} & \text{if } j \to i \text{ in } C \\ 0 & \text{otherwise} \end{cases} \tag{49}$$

for each n-cycle C in the decomposition of G. Then the collection of matrices $\{X_C, X_{-C}\}_{C \subseteq G}$ provides a means of estimating the off-diagonal entries of the outer product $\omega\omega^*$.

Let A denote the adjacency matrix of G and consider the subspace T of the vector space of self-adjoint $n \times n$ matrices $\mathbb{H}^{n \times n}$ defined by

$$T := \{X \in \mathbb{H}^{n \times n} : X[i,j] = 0 \text{ whenever } A[i,j] = 1\}.$$

Notice that the nonzero entries of any matrix $X \in T$ do not coincide with those of any matrix in the collection $\{X_C, X_{-C}\}_{C \subseteq G}$. Thus, defining the Minkowski sum $S := \operatorname{span}\{X_C, X_{-C}\}_{C \subseteq G} + T$, it follows that $\omega\omega^* \in S$ (for the noiseless case).

Finally, since the diagonal entries of $\omega\omega^*$ are all equal to 1, defining

$$\delta_i(x) := \begin{cases} 1 & \text{if } x = i \\ 0 & \text{otherwise.} \end{cases}$$

for each $i \in \{0, \dots, n-1\}$ yields $\langle \omega\omega^*, \delta_i\delta_i^* \rangle_{\mathrm{HS}} = 1$ for every $i = 0, \dots, n-1$. Combining this with the above result, we see that one may obtain an estimate for the outer product $\omega\omega^*$ by seeking a rank-1 matrix as close to the subspace S as possible whose diagonal entries are nearly 1. This leads to the following feasibility problem for some fixed tolerance $\varepsilon > 0$; here, $\mathcal{P}_V X$ denotes the projection of X onto the subspace V:

$$\begin{aligned} \text{Find} \quad X \in \mathbb{H}^{n\times n} \qquad \text{such that} \quad & \operatorname{rank}(X) = 1 \\ \text{and} \quad & \|\mathcal{P}_{S^\perp} X\|_{\mathrm{HS}}^2 + \sum_{i=0}^{n-1} \big| 1 - \langle X, \delta_i\delta_i^* \rangle_{\mathrm{HS}} \big|^2 \leq \varepsilon. \end{aligned} \tag{50}$$

Due to the tolerance $\varepsilon > 0$, we expect that the above program would be particularly stable to noise on the order of ε. Unfortunately, this program is not convex and so is not easily solved. To make it convex, one could relax the rank-1 condition to a positive-semidefinite requirement, but this makes the feasibility set too large; indeed, the identity matrix is necessarily feasible for the relaxed problem but has full rank. Properly relaxing the problem remains an open problem. Note that the outer product of any modulation of ω is feasible in (50). As such, any convex combination of these will be feasible in a convex relaxation. In Appendix B, we show how to determine ω up to a modulation and a global phase factor from any such convex combination.

5.4 Solving the phase error problem in the noisy case

The process of determining phase errors from nested relative phases via the feasibility problem (50) of the previous section appears to be quite difficult. For this reason, we seek the phase errors by another method. In particular, we will obtain them by applying angular synchronization iteratively.

Let $\omega = \{\omega_i\}_{i\in V} \subseteq \mathbb{T}$ be a vector of unknown phase errors and consider the graphs $G = (V, E)$ and $G' = (V', E')$ from the previous section, where G is circulant with n vertices such that n is an odd prime. Thus, we encode the unknown relative phases as edge measurements in G, namely $\{\omega_i\omega_j^{-1}\}_{((i,j))\in E}$, and similarly for the known (noisy) relative phases in G', $\{\sigma_{i,j}\sigma_{j,k}^{-1} + \varepsilon_{(i,j),(j,k)}\}_{((i,j),(j,k))\in E'}$, where $\sigma_{i,j} := \omega_i\omega_j^{-1}$ for all $i, j \in \{0, \ldots, n-1\}$ and each $\varepsilon_{(i,j),(j,k)}$ is an adversarial noise term (under the assumption that $\overline{\varepsilon_{(i,j),(j,k)}} = \varepsilon_{(k,j),(j,i)}$). Since Algorithm 3 (angular synchronization) requires the input graph be directed, suppose we arbitrarily direct the edges of G and G' such that edges in a common n-cycle have the same orientation. To be consistent, this means we take only the relative phases $\omega_i\omega_j^{-1}$ if $(i,j) \in E$ or $\omega_j\omega_i^{-1} = \overline{\omega_i\omega_j^{-1}}$ if $(j,i) \in E$ (likewise for $\sigma_{i,j}\sigma_{j,k}^{-1} + \varepsilon_{(i,j),(j,k)}$ as edges in E').

Unfortunately, the structure G' inherits from G is not conducive to angular synchronization. In fact, each n-cycle in the decomposition of G corresponds to a distinct n-cycle in G', each of which forming its own component. Thus, G' itself is not connected and, therefore, has a spectral gap of 0. For this reason, it doesn't even make sense to perform angular synchronization on G'. We can, however, use the cycles in G', along with the corresponding edge measurements, as separate inputs for Algorithm 3. Since the vertices in G' correspond to edges in G, the outputs then estimate certain entries of the phase error outer product $\omega\omega^*$, specifically those represented by the edges of each cycle in G. By Lemma 5.8, it follows that the information encoded on the edges of distinct cycles in G are obtained up to their own global phase factors. Assuming one could "synchronize" these phases in such a way that each cycle has the same global phase, it would then be possible to use G and the relative phases $\{\omega_i\omega_j^{-1}\}_{((i,j))\in E}$ as inputs in Algorithm 3 to obtain an estimate for the phase error vector ω. Admittedly, this is the most ad hoc portion of the phase error recovery process; indeed, the feasibility problem (50) was considered as a possible way to circumvent this issue. The problem with this intermediate step is that the global phase factors associated with each cycle in G interact differently depending on the cycle orientations. To see this, note that each n-cycle in G necessarily contains every vertex of G, and so traversing a given cycle along adjacent edges generates a unique walk in G which reaches every vertex exactly once before repeating. Because each cycle is distinct, however, the number of steps necessary to reach a certain vertex from the same starting vertex in any two walks is different. Recalling that phase propagates along edges, this means that the accumulated phase due to the global phase factors varies from cycle to cycle. Overall, it would be much better to estimate the relative phases $\{\omega_i\omega_j^{-1}\}_{(i,j)\in E}$ from $\{\sigma_{i,j}\sigma_{j,k}^{-1}\}_{((i,j),(l,k))\in E'}$ in one step instead of synchronizing components of G individually before synchronizing their outputs. In the absence of a better alternative, we continue.

To synchronize the cycles, we take advantage of the circulant property of the adjacency matrices of each cycle in G; in particular, note that it is possible to obtain common support amongst these matrices simply by raising each to a particular non-negative power. These exponents represent the different number of steps necessary to move in each cycle in order to reach the same vertex given a common starting vertex. Since cycles must exhibit consistency in relative phase (i.e., the product of relative phase between any two vertices is constant regardless of the cycle chosen), raising each matrix to the appropriate power without the global phase factors should yield a set of identical circulant matrices. However, the fact that the exponents differ means that the global phase factors are not related linearly. For instance, to move between adjacent vertices in one cycle may require two steps in another cycle, and so their corresponding phase factors have a quadratic relationship; others may yet have even higher order. To determine each cycle up to a common global phase, it is therefore necessary to incorporate all of these phase relationships.

Let $S \subseteq \mathbb{N}$ denote a set of indices such that each distinct n-cycle in G is labeled C_i for some $i \in S$. Furthermore, assume that two n-cycles C_i and C_j in G are distinct if and only if $i, j \in S$ and $i \neq j$. Consider the function $s\colon S \times S \to \mathbb{Z}_n$ which outputs the exponent such that X_{C_i} and $X_{C_j}^{s(i,j)}$ have common support; here, X_{C_i} denotes the weighted adjacency matrix of the cycle C_i defined in (49). Then by the above discussion there exists a function $\alpha\colon S \times S \to \mathbb{T}$ such that

$$\alpha(i,j)X_{C_i} = X_{C_j}^{s(i,j)}$$

for all $i, j \in S$. Since these weighted adjacency matrices are related to the actual adjacency matrices of the cycles by distinct global phase factors, denoting the phase factor associated with the cycle C_i by β_i then yields the relation

$$\frac{\alpha(i,j)}{\beta_i} = \Big(\frac{1}{\beta_j}\Big)^{s(i,j)}$$

for all $i, j \in S$. Rearranging, we see that the phases $\{\beta_i\}_{i \in S}$ may be expressed as $\beta_i = \alpha(i,j)\beta_j^{s(i,j)}$ for all $i, j \in S$. Thus, we require an initial guess for one of the phases in order to generate the entire set. Such a guess is not an issue, since it merely assigns an arbitrary phase to one of the cycles in G, in terms of which the remaining phase factors are determined. To be clear, we already know there must be an ambiguity in the result regarding a single global phase factor, making this step legitimate. Rather than simply choosing a random phase to generate the set of phases, it makes sense to choose a particular β_i such that it scales X_{C_i} to have phase 1; i.e., we choose β_i^{-n} to be equal to the product of the nonzero entries of X_{C_i} for some fixed $i \in S$. Using this fixed phase to generate the remaining phases then yields the set $\{\beta_i\}_{i \in S}$ up to a single global phase; we outline this process in its entirety in Algorithm 4. As we saw in Lemma 5.8 and Corollary 5.9, the consequence of this global phase is to ultimately cause the estimated phase error vector to be a modulation of the true phase error vector.

Now that we have recovered the phase error vector up to a global phase and modulation, a SAR image may be reconstructed using a simple algorithm, which we now discuss. Recall that the phase error vector provides the phase factors associated with slices of the Fourier transform of the target image obtained according to Fact 5.2. To recover the modulation, we leverage the effect that modulating the

Algorithm 4 Cycle synchronization

```
Input: Weighted $n \times n$ adjacency matrices $\{X_{C_i}\}_{i \in S}$ defined in (49)
Output: Vector $\{\beta_i\}_{i \in S}$ of cycle-dependent phase factors
  Initialize $|S| \times |S|$ matrices $s$ and $\alpha$ of all zeros
  for $k, j = 1$ to $|S|$ do
    Compute $p \in \mathbb{Z}_n$ and $c \in \mathbb{T}$ such that $cX_{C_k} = X_{C_j}^p$
    Assign $s[k, j] \leftarrow p$ and $\alpha[k, j] \leftarrow c$
  end for
  Fix $\ell \in S$
  Compute $\beta_\ell \leftarrow \left(\prod_{k \to j} X_{C_\ell}[k, j]\right)^{-1/n}$    {Initialize $\beta_\ell$ in terms of the product of all nonzero entries of $X_{C_\ell}$}
  for $k = 1$ to $|S|$, $k \neq \ell$ do
    Compute $\beta_k \leftarrow \alpha[k, \ell]\beta_\ell^{s[k,\ell]}$
  end for
  Output: $\{\beta_i\}_{i \in S}$
```

slices in the Fourier domain has on the *total variation* of the image in the spatial domain. The total variation of an image is the sum of the absolute values of the differences between adjacent pixels in the image. Since a modulation of the slices in the Fourier domain is a pointwise multiplication of the Fourier transform, it has the effect of a convolution in the spatial domain with some function, essentially blurring the image. Thus, sharp contrasts in the image are reduced; in other words, those adjacent pixels which contribute more to the total variation are changed in such a way that their contribution is reduced, and so a modulation in the Fourier domain reduces the total variation in the spatial domain. Consequently, the best estimate for the true modulation is that which maximizes the total variation of the image. In order to be independent of the unknown global phase factor, we actually take the modulation which maximizes the total variation of the *absolute value* of the image.

As for the global phase factor, suppose the true image data in the spatial domain is all real and nonnegative. Then taking the two dimensional inverse Fourier transform of the data with the estimated modulation determined above will yield values in the spatial domain that are (approximately) scaled versions of the global phase. Thus, adding all pixel values and normalizing gives an estimate for the global

Algorithm 5 Total variation maximization

```
Input: Slices of Fourier transform data, associated phase errors $\omega := \{\omega_i\}_{i=0}^{n-1}$ up to a modulation and global phase factor
Output: Estimated SAR image $\hat{Y}$
  Let $X$ denote the matrix of Fourier transform data
  Let $M$ denote the matrix of slice indices $\{0, \ldots, n-1\}$
  for $k = 0$ to $n-1$ do
    Compute $\widetilde{\omega} \leftarrow E^k \omega$
    Compute $\hat{X} \leftarrow X \circ \widetilde{\omega}(M)$          {$\circ$ is the Hadamard matrix product}
    Compute the inverse Fourier transform $Y \leftarrow F^{-1}\hat{X}$
    Compute the total variation $TV_k \leftarrow \|Y\|_{TV}$
  end for
  Compute $m \leftarrow \arg\max_k TV_k$
  Compute $\widetilde{\omega} \leftarrow E^m \omega$
  Compute $\hat{X} \leftarrow X \circ \widetilde{\omega}(M)$
  Compute the inverse Fourier transform $Y \leftarrow F^{-1}\hat{X}$
  Compute $\psi \leftarrow \sum_i \sum_j Y[i,j] / |\sum_i \sum_j Y[i,j]|$
  Output: $\hat{Y} \leftarrow \mathrm{Re}(\psi^{-1} Y)$
```

phase which, combined with the estimated modulation in the Fourier domain, will yield the best estimate for the image itself (after an inverse Fourier transform). Due to the effect of noise, we take the real part of the resulting image data to approximate the actual image, since the imaginary parts of the pixel data are expected to be relatively small. This procedure is outlined in Algorithm 5.

Overall, the solution to the phase error problem we have developed here involves a particular measurement design and phase error reconstruction procedure:

Measurement design

- Let $G = (V, E)$ and $G' = (V', E')$ be graphs such that G is circulant with $|V|$ an odd prime, $V' = E$, and $((i,j),(k,\ell)) \in V'$ are adjacent in G' if and only if they share a neighbor in a common n-cycle in the decomposition of G given by Lemma 5.7

- Design a multistatic SAR system such that applying Fact 5.2 yields the unknown phase errors $\{\omega_i\}_{i\in V}$, the unknown relative phases $\{\omega_i\omega_j^{-1}\}_{(i,j)\in V'}$, and the known combinations $\{\omega_i\omega_j^{-2}\omega_k\}_{((i,j),(j,k))\in E'}$

SAR image reconstruction procedure

- For each cycle in G', use Algorithm 3 to calculate the relative phases $\{\omega_i\omega_j^{-1}\}_{(i,j)\in C\subseteq G'}$ up to distinct global phase factors from $\{\omega_i\omega_j^{-2}\omega_k\}_{((i,j),(j,k))\in E'}$
- Form the weighted adjacency matrices for each n-cycle in G using the relative phases $\{\omega_i\omega_j^{-1}\}_{(i,j)\in G'}$ from the corresponding cycles in G' according to (49)
- Use Algorithm 4 to synchronize the cycle-dependent phase factors
- Use Algorithm 3 to calculate the phase errors $\{\omega_i\}_{i\in V}$ up to a modulation and a global phase factor from the synchronized relative phases $\{\omega_i\omega_j^{-1}\}_{(i,j)\in V'}$
- Use Algorithm 5 to estimate the SAR image by picking the modulation that maximizes total variation

To test the phase error recovery procedure (i.e., the first four bullets of the SAR image reconstruction procedure above), we simulated the recovery of random phase errors ω_i from noisy products of the form $\omega_i\omega_j^{-2}\omega_k$ using random circulant graphs on 101 vertices. For graphs G containing a fixed number of cycles, we generated a random phase error vector and took such a product of phases for each edge in the corresponding graph G'. We then added complex Gaussian noise to each product (normalizing the resultant entries) to simulate noisy SAR data acquired from multistatic schemes according to Fact 5.2. Finally, using Algorithms 3 and 4 as prescribed above, we generated the estimated phase error vector (up to a modulation and global phase factor). Figure 6 depicts how the performance of this phase error reconstruction changed with the number of cycles in the graph G. Since the global phase ambiguity is not resolved in this portion of the image reconstruction process, the relative error here is measured between the outer products of the phase error vector and the phase error estimate, which we then minimized over all possible

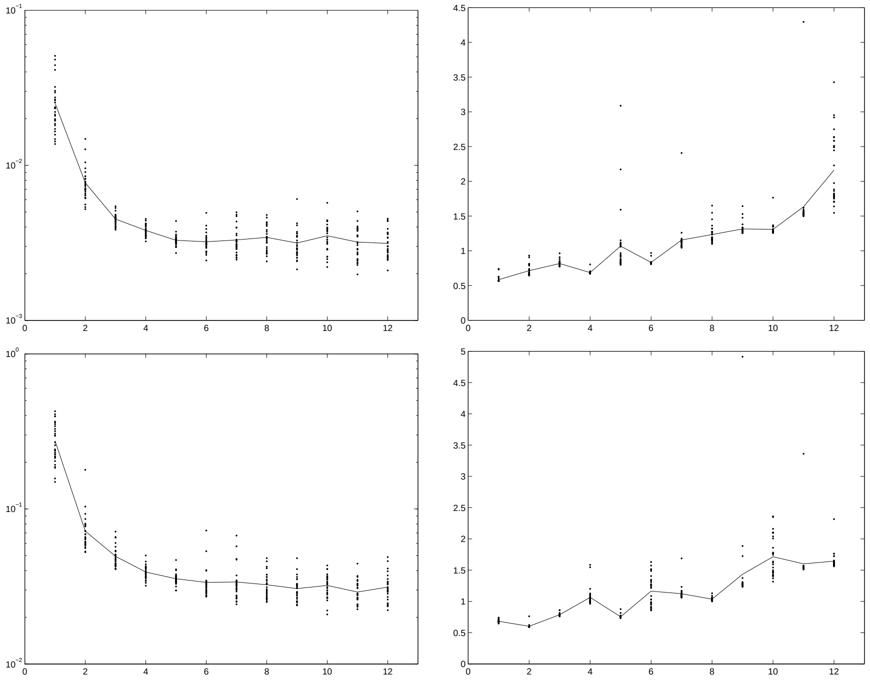


Figure 6: Iterative angular synchronization as a means of recovering a phase error vector (up to a modulation and global phase) from nested relative phase information. Here, simulations use a graph with 101 vertices and varying numbers of cycles. Note that the global phase ambiguity in the estimate cannot be resolved, and so the relative errors are taken between the outer products of the true and estimated phase error vectors. Since the estimate produced is also a modulation of the true phase error vector, the errors depicted are the minima over all possible modulations. At left we plot relative error in the output as a function of the number of cycles in G, while at right we plot computation time for the same varying set of cycles. For the first row of plots (top), the normalized noise is fixed in average magnitude at $\sigma^2 = 0.01$, while the second row (bottom) depicts noise fixed at $\sigma^2 = 0.1$. Piecewise linear graphs connecting the sample averages are shown for clarity.

modulations. Note that multiple cycles tend to reduce the error in reconstruction, though with diminishing returns. Since the computation time grows with each cycle added to G, this suggests the existence of an “optimal” number of cycles that gives good reconstruction while keeping computation time low.

To get an idea of the behavior of the phase error recovery procedure in the presence of noise, we also simulated the recovery of random phase errors from noisy multistatic SAR data in different noise regimes using a fixed number of cycles. In particular, we generated random circulant graphs containing 1, 5, and 10 cycles, examining the performance of phase error recovery in each case under varying levels of noise. Figure 7 displays the observed relative error in output as a function of

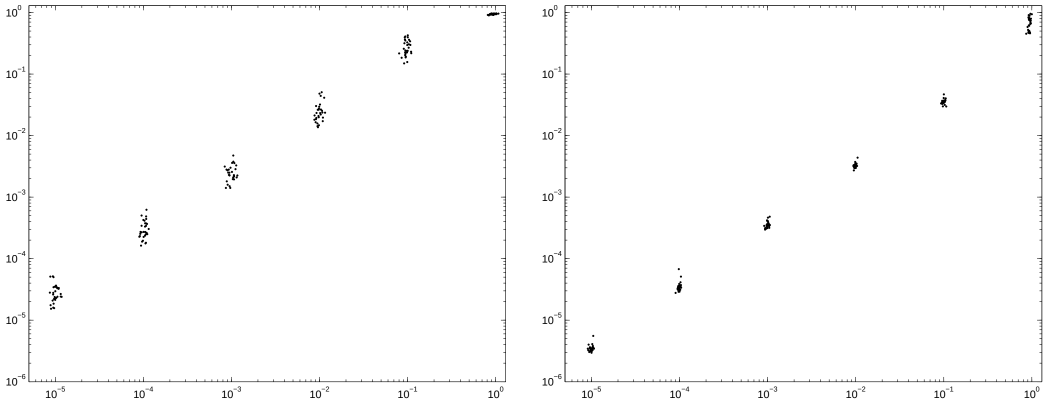


Figure 7: The stability of iterative angular synchronization is cycle-dependent—more cycles imply better stability (and each cycle comes from a pair of aircraft in the multistatic system). Here, iterative angular synchronization recovers the phase error vector (up to a modulation and global phase) from simulated nested relative phase information using a graph with 101 vertices and varying numbers of cycles. Since the estimate produced is necessarily a modulation of the true phase error vector, the errors depicted are taken to be the minima over all possible modulations. Each plot depicted shows relative error in recovery as a function of average input noise magnitude (normalized). From left to right, the graphs used contain 1 and 5 cycles; the case of 10 cycles behaves almost identically to the 5 cycle case. Note when the graph contains one cycle the algorithm is relatively ineffective at determg the phase error vector (up to any modulation) regardless of the level of noise in the model; as such stability is poor for this case. Meanwhile, graphs with greater numbers of cycles exhibit better stability to noise.

input noise; as illustrated, the phase error reconstruction procedure appears to be stable to noise provided the graph G contains multiple cycles.

Finally, we tested the last bullet of the SAR image reconstruction procedure. To do this, we used a cropped version of a SAR image of the Pentagon (available at [83]). Partitioning the pixels of its Fourier transform into 101 sections, each corresponding to a different slice of Fourier space, we then multiplied each section by a different (noisy) phase factor to simulate phase errors. Specifically, we generated a vector of 101 iid random phases and added complex Gaussian noise, normalizing each resultant entry. Since the phase error recovery procedure only guarantees recovery of the true phase errors up to a modulation and global phase factor, we then randomly modulated this vector and multiplied by a random phase; pointwise multiplying each entry of this vector by the corresponding slice of the image's Fourier transform then simulates the output of the first four bullets of the SAR image reconstruction procedure. Taking the actual phase error vector to be all ones (without loss of

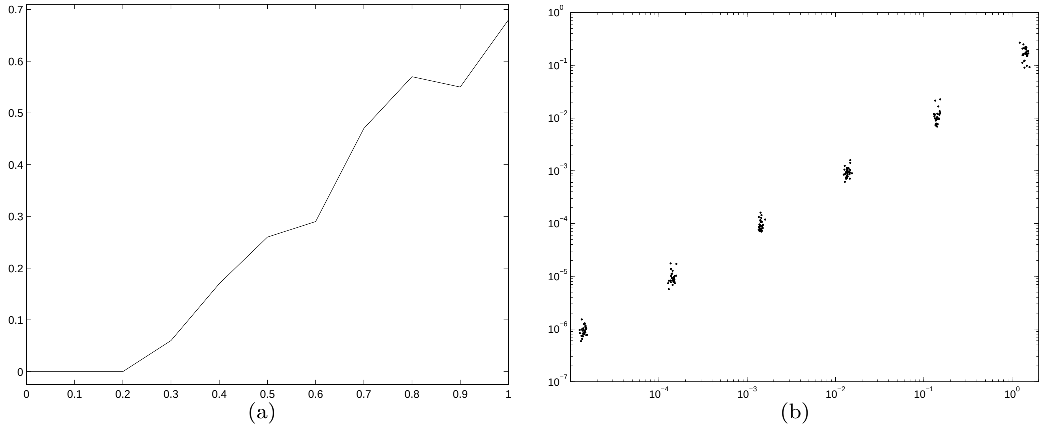


Figure 8: The Fourier domain is partitioned into 101 lines and each lines is given a simulated (noisy) phase error. If the phase errors are known up to a modulation and global phase factor, then Algorithm 5 recovers the modulation by maximizing total variation of the image. (a) The (approximate) probability that the estimated modulation is not the true modulation (as a function of average noise magnitude); here, a sample of 100 simulations provide the estimated probability for each noise level in $\{0, 0.1, 0.2, \ldots, 1\}$. (b) The relative error in image reconstruction using phase errors estimated by Algorithm 5 suggests stability to noise; depicted are clusters of 30 simulations at varying noise levels.

generality), we then gave the generated phase error estimates and noisy Fourier data to Algorithm 5, the results of which are depicted in Figures 8 and 9.

As we can see, the demodulation portion of this algorithm is particularly stable to noise; indeed, Figure 8(a) suggests that the total variation method of modulation recovery has a success rate of over 90% for noise with average magnitude of up to 0.3. Also, Figure 8(b) shows that relative error in image reconstruction behaves well with the input noise. Figure 9 illustrates how the total variation of the image under all possible modulations is affected by differing levels of noise, as well as the resultant quality in image recovery. Indeed, the noiseless case demonstrates that the true modulation is easily implicated by a large total variation. When the noise level is 0.3 in average magnitude, the true modulations can still be detected by maximizing total variation, but the sidelobes are rising and the noise is somewhat noticeable in the resulting image reconstruction. Finally, using noise with average magnitude of 0.7 raises the sidelobes so high that the true modulation can no longer be detected. As Figure 8(a) indicates, we can expect the true modulation to be successfully detected only 50% of the time with this level of noise.

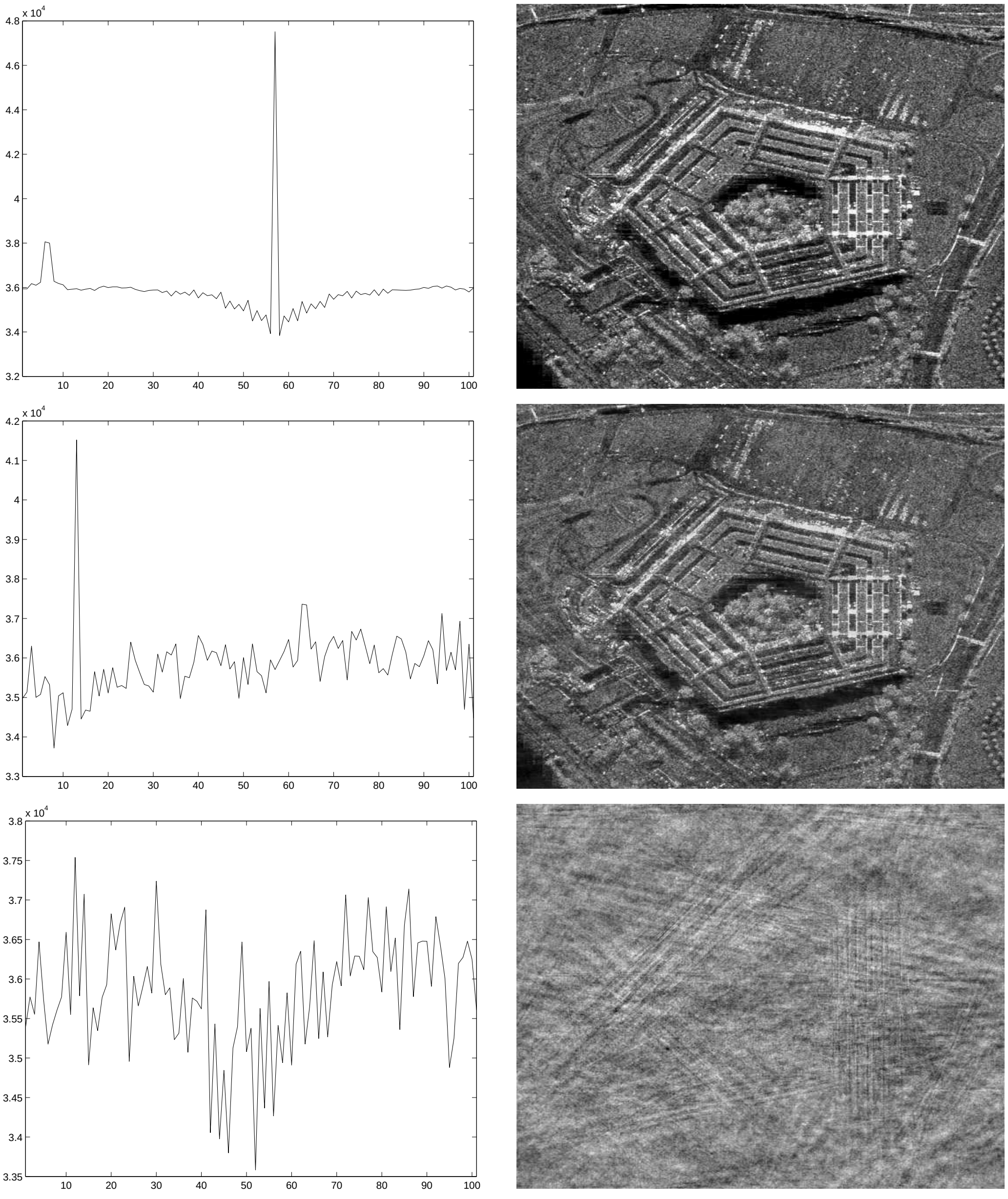


Figure 9: Reconstructing a SAR image from noisy Fourier data. The Fourier domain is partitioned into 101 lines and the image information on each line is given a random (noisy) global phase. The resultant simulated phase error vector is then modulated with a random modulation index. To recover the modulation, the image is reconstructed using all possible modulations and that which yields maximal total variation in the reconstruction is taken; the global phase factor is then estimated to reconstruct the image with this choice of modulation (see Algorithm 5). At top right is a SAR image of the Pentagon (noiseless); the images beneath are reconstructions with average noise magnitudes 0.3 and 0.7, respectively. Each plot at left shows the total variation of the corresponding image as a function of modulation index. For the noiseless case, the maximizing modulation 57 is indeed the true modulation. With noise at average magnitude of 0.3, the total variation is again maximized at the true modulation index (13). In the case of average noise magnitude 0.7, the modulation which maximizes the total variation is 12, while the true modulation is 45, resulting in the unsuccessful recovery displayed. The relative error in the reconstructed images at noise levels of 0.3 and 0.7 above are 0.0410 and 0.1667, respectively.

VI. Conclusion

In this thesis we made significant theoretical progress on the phase retrieval problem. First, we analyzed what it means for an ensemble of intensity measurements to be injective and stable. In doing so, we characterized injectivity in both the real and complex cases, leading to the conjecture that $4M-4$ intensity measurements are necessary and sufficient to successfully determine an M-dimensional complex signal (up to a global phase). We made certain contributions toward a proof of this conjecture, to include a deterministic construction of an injective ensemble using $4M-4$ intensity measurements. Next, we devised a theory of almost injectivity to characterize ensembles of intensity measurements that enable signal recovery on a dense subset of $\mathbb{C}^M$. This led to a discussion of the computational limits of phase retrieval, where we analyzed computational complexity by drawing connections between phase retrieval and the well-known subset sum problem. We concluded with a stability analysis, where we developed stronger versions of the characterizations of injectivity in the presence of stochastic noise as well as a new condition which strengthens the complement property in the worst case.

The second major contribution of this thesis was to develop a new multistatic methodology for synthetic aperture radar to resolve phase errors. Drawing motivation from the phase retrieval problem, we related the phase error problem to interferometric phase recovery techniques. Using graphs to organize the SAR data, we then leveraged angular synchronization to reconstruct the phase errors. Simulations suggest that image reconstruction based on this approach is stable to noise, and desirable results can be achieved using few aircraft.

Aside from this thesis, theoretical progress on the phase retrieval problem is currently an active area of research. For instance, the $4M-4$ Conjecture, although widely believed, still lacks a complete proof. The following summarizes what is currently known about the conjecture:

- The conjecture holds for $M = 2$ and $M = 2^m + 1$, $m = 1, 2, 3, \dots$ [38] (see also Section 2.1 of this thesis).
- If $N < 4M - 2\alpha(M-1) - 3$, then $\mathcal{A}$ is not injective [62]; here, $\alpha(M-1) \leq \log_2 M$ denotes the number of 1's in the binary expansion of $M - 1$.
- For each $M \geq 2$, there exists an ensemble Φ of $N = 4M - 4$ measurement vectors such that $\mathcal{A}$ is injective [17] (see also Section 2.3 of this thesis).
- If $N \geq 4M - 4$, then $\mathcal{A}$ is injective for generic Φ [38] (cf. [8]).

Furthermore, there remains a fundamental lack of understanding of injectivity in the complex case. Although injectivity has been characterized in this case (see Sections 2.1 and 2.2 of this thesis), formulating sufficient conditions for injectivity in the complex case which can be verified in finite time remains a subject for future work. Characterizing almost injectivity for complex ensembles is also an open problem, and a stability analysis for phase retrieval in this setting is still necessary.

The phase error problem in synthetic aperture radar similarly requires future work. In particular, alternative ways of relaxing the feasibility problem (50) to enable the use of convex programming techniques is highly desirable. If this is not possible, the process described in Section 5.4 requires a performance guarantee, and it may also be improved upon. Indeed, the intermediate step of cycle synchronization to link the two graphs G and G' is not very democratic; an alternative method that simultaneously uses information from all available cycles in G' might improve efficiency and stability of the entire algorithm. Meanwhile, the phase error recovery algorithm presented in this thesis is in need of a feasibility assessment to better understand the possibility of real-world implementation in SAR imaging systems. For example, what types of (realistic) aircraft flight patterns enable the encoding of multistatic SAR data using circulant graphs? Also, are there other types of graphs, which may be more easily implementable, that exhibit the same phase error recovery characteristics? In this thesis, we assumed that we knew when an aircraft's bearing vector bisects the bearing vectors of two other aircraft—how stable is this assumption to fluctuations in the aircraft position? Further issues concerning the effect of crosstalk in multistatic SAR systems need to be addressed as well, and questions regarding simplifying assumptions should be considered (e.g., two-dimensional target scenes). One should account for all of these factors, not only in discerning the feasibility of the multistatic methodology we present, but also in competing with state of the art phase error–correction algorithms, which do not require a multistatic system but might prove to be less reliable.

Appendix A.

Here, we verify that we can differentiate under the integral sign in the proof of Theorem 4.10 from Section 4.2.

Lemma A.1. *Consider the probability density function defined by*

$$f(y;\theta) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\|y-\mathcal{A}(\theta)\|^2/2\sigma^2} \qquad \forall y \in \mathbb{R}^N.$$

Then for every function $g\colon \mathbb{R}^N \to \mathbb{R}$ *with finite second moment*

$$\int_{\mathbb{R}^N} g(y)^2 f(y;\theta)dy < \infty \qquad \forall \theta \in \Omega,$$

we can differentiate under the integral sign:

$$\frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^N} g(y) f(y;\theta)dy = \int_{\mathbb{R}^N} g(y)\frac{\partial}{\partial\theta_i} f(y;\theta)dy.$$

Proof. First, we adapt the proof of Lemma 5.14 in [72] to show that it suffices to find a function $b(y;\theta)$ with finite second moment such that, for some $\varepsilon > 0$,

$$\left|\frac{f(y;\theta + z\delta_i) - f(y;\theta)}{zf(y;\theta)}\right| \leq b(y;\theta) \qquad \forall y \in \mathbb{R}^N, \theta \in \Omega, |z| < \varepsilon, z \neq 0 \tag{51}$$

where δ_i denotes the ith identity basis element in $\mathbb{R}^{2M}$. Indeed, by applying the Cauchy-Schwarz inequality over f-weighted L^2 space, we have

$$\begin{aligned}&\int_{\mathbb{R}^N} |g(y)| b(y;\theta) f(y;\theta)dy \\ &\qquad\leq \left(\int_{\mathbb{R}^N} g(y)^2 f(y;\theta)dy\right)^{1/2}\left(\int_{\mathbb{R}^N} b(y;\theta)^2 f(y;\theta)dy\right)^{1/2} < \infty\end{aligned}$$

and so the dominated convergence theorem gives

$$
\begin{aligned}
\int_{\mathbb{R}^N} g(y)\frac{\partial}{\partial\theta_i}f(y;\theta)dy &= \int_{\mathbb{R}^N}\lim_{z\to 0}\left(g(y)\frac{f(y;\theta+z\delta_i)-f(y;\theta)}{zf(y;\theta)}\right)f(y;\theta)dy \\
&= \lim_{z\to 0}\int_{\mathbb{R}^N}\left(g(y)\frac{f(y;\theta+z\delta_i)-f(y;\theta)}{zf(y;\theta)}\right)f(y;\theta)dy \\
&= \lim_{z\to 0}\frac{1}{z}\left(\int_{\mathbb{R}^N} g(y)f(y;\theta+z\delta_i)dy - \int_{\mathbb{R}^N} g(y)f(y;\theta)dy\right) \\
&= \frac{\partial}{\partial\theta_i}\int_{\mathbb{R}^N} g(y)f(y;\theta)dy.
\end{aligned}
$$

In pursuit of such a function $b(y;\theta)$, we first use the triangle and Cauchy-Schwarz inequalities to get

$$
\begin{aligned}
\left|\frac{f(y;\theta+z\delta_i)-f(y;\theta)}{zf(y;\theta)}\right| &= \frac{1}{|z|}\left|e^{-\frac{1}{2\sigma^2}\left(\|y-\mathcal{A}(\theta+z\delta_i)\|^2-\|y-\mathcal{A}(\theta)\|^2\right)}-1\right| \\
&= \frac{1}{|z|}\Big|e^{-\frac{1}{2\sigma^2}\left(\|\mathcal{A}(\theta+z\delta_i)\|^2-\|\mathcal{A}(\theta)\|^2-2\langle y,\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\rangle\right)} \\
&\qquad - e^{\frac{1}{\sigma^2}\langle y,\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\rangle} + e^{\frac{1}{\sigma^2}\langle y,\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\rangle} - 1\Big| \\
&\le \frac{1}{|z|}\Bigg(e^{\frac{1}{\sigma^2}\langle y,\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\rangle}\left|e^{-\frac{1}{2\sigma^2}\left(\|\mathcal{A}(\theta+z\delta_i)\|^2-\|\mathcal{A}(\theta)\|^2\right)}-1\right| \\
&\qquad + \left|e^{\frac{1}{\sigma^2}\langle y,\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\rangle}-1\right|\Bigg) \\
&\le \frac{1}{|z|}\Bigg(e^{\frac{1}{\sigma^2}\|y\|\|\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\|}\left|e^{-\frac{1}{2\sigma^2}\left(\|\mathcal{A}(\theta+z\delta_i)\|^2-\|\mathcal{A}(\theta)\|^2\right)}-1\right| \\
&\qquad + \left|e^{\frac{1}{\sigma^2}\|y\|\|\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\|}-1\right|\Bigg), \qquad (52)
\end{aligned}
$$

Denote $c(z;\theta) := \frac{1}{\sigma^2}\|\mathcal{A}(\theta+z\delta_i)-\mathcal{A}(\theta)\|$. Since $(e^{st}-1)/t \le se^{st}$ whenever $s,t\ge 0$, we then have

$$
\frac{|e^{c(z;\theta)\|y\|}-1|}{|z|} = \frac{c(z;\theta)}{|z|}\cdot\frac{e^{c(z;\theta)\|y\|}-1}{c(z;\theta)} \le \frac{c(z;\theta)}{|z|}\|y\|e^{c(z;\theta)\|y\|}.
$$

Also by l'Hospital's rule, there exist continuous functions C_1 and C_2 on the real line such that

$$C_1(z;\theta) = \frac{c(z;\theta)}{|z|}, \qquad C_2(z;\theta) = \frac{1}{|z|}\left|e^{-\frac{1}{2\sigma^2}\left(\|\mathcal{A}(\theta+z\delta_i)\|^2 - \|\mathcal{A}(\theta)\|^2\right)} - 1\right|, \qquad \forall z \neq 0.$$

Thus, continuing (52) gives

$$\left|\frac{f(y;\theta+z\delta_i) - f(y;\theta)}{zf(y;\theta)}\right| \leq \Big(C_1(z;\theta)\|y\| + C_2(z;\theta)\Big)e^{c(z;\theta)\|y\|}.$$

Now for a fixed ε, take $C_j(\theta) := \sup_{|z|<\varepsilon} C_j(z;\theta)$ and $c(\theta) := \sup_{|z|<\varepsilon} c(z;\theta)$, and define

$$b(y;\theta) := \Big(C_1(\theta)\|y\| + C_2(\theta)\Big)e^{c(\theta)\|y\|}.$$

Since $C_j(\theta)$ and $c(\theta)$ are suprema of continuous functions over a bounded set, these are necessarily finite for all $\theta \in \Omega$. As such, this choice for b satisfies (51). It remains to verify that b has a finite second moment. To this end, let $B(R(\theta))$ denote the ball of radius $R(\theta)$ centered at the origin (we will specify $R(\theta)$ later). Then

$$\begin{aligned}\int_{\mathbb{R}^N} b(y;\theta)^2 f(y;\theta)dy &= \int_{B(R(\theta))} b(y;\theta)^2 f(y;\theta)dy + \int_{\mathbb{R}^N\setminus B(R(\theta))} b(y;\theta)^2 f(y;\theta)dy \\ &\leq \Big(C_1(\theta)R(\theta) + C_2(\theta)\Big)^2 e^{2c(\theta)R(\theta)} \\ &\quad + \frac{1}{(2\pi\sigma^2)^{N/2}}\int_{\mathbb{R}^N\setminus B(R(\theta))} \Big(C_1(\theta)\|y\| + C_2(\theta)\Big)^2 e^{2c(\theta)\|y\| - \frac{1}{2\sigma^2}\|y-\mathcal{A}(\theta)\|^2}dy.\end{aligned} \tag{53}$$

From here, we note that whenever $\|y\| \geq 2\|\mathcal{A}(\theta)\| + 8\sigma^2 c(\theta)$, we have

$$\begin{aligned}\|y-\mathcal{A}(\theta)\|^2 &\geq \|y\|^2 - 2\|y\|\|\mathcal{A}(\theta)\| + \|\mathcal{A}(\theta)\|^2 \\ &\geq \Big(2\|\mathcal{A}(\theta)\| + 8\sigma^2 c(\theta)\Big)\|y\| - 2\|y\|\|\mathcal{A}(\theta)\| + \|\mathcal{A}(\theta)\|^2 \\ &\geq 8\sigma^2 c(\theta)\|y\|.\end{aligned}$$

Rearranging then gives $2c(\theta)\|y\| \leq \frac{1}{4\sigma^2}\|y - \mathcal{A}(\theta)\|^2$. Also let $h(\theta)$ denote the larger root of the polynomial

$$p(x;\theta) := 2C_1(\theta)^2\Big(x^2 - 2\|\mathcal{A}(\theta)\|x + \|\mathcal{A}(\theta)\|^2\Big) - \Big(C_1(\theta)x + C_2(\theta)\Big)^2,$$

and take $h(\theta) := 0$ when the roots of $p(x;\theta)$ are not real. (Here, we are assuming that $C_1 > 0$, but the proof that (53) is finite when $C_1 = 0$ quickly follows from the $C_1 > 0$ case.) Then $(C_1(\theta)\|y\| + C_2(\theta))^2 \leq 2C_1(\theta)^2\|y - \mathcal{A}(\theta)\|^2$ whenever $\|y\| \geq h(\theta)$, since by the Cauchy-Schwarz inequality,

$$2C_1(\theta)^2\|y - \mathcal{A}(\theta)\|^2 - \Big(C_1(\theta)\|y\| + C_2(\theta)\Big)^2 \geq p(\|y\|;\theta) \geq 0,$$

where the last step follows from the fact that $p(x;\theta)$ is concave up. Now we continue by taking $R(\theta) := \max\{2\|\mathcal{A}(\theta)\| + 8\sigma^2 c(\theta), h(\theta)\}$:

$$\begin{aligned}
&\int_{\mathbb{R}^N \setminus B(R(\theta))} \Big(C_1(\theta)\|y\| + C_2(\theta)\Big)^2 e^{2c(\theta)\|y\| - \frac{1}{2\sigma^2}\|y - \mathcal{A}(\theta)\|^2} dy \\
&\quad \leq \int_{\mathbb{R}^N \setminus B(R(\theta))} 2C_1(\theta)^2\|y - \mathcal{A}(\theta)\|^2 e^{-\frac{1}{4\sigma^2}\|y - \mathcal{A}(\theta)\|^2} dy \\
&\quad \leq \big(2\pi(\sqrt{2}\sigma)^2\big)^{N/2} \cdot 2C_1(\theta)^2 \int_{\mathbb{R}^N} \|x\|^2 \frac{1}{(2\pi(\sqrt{2}\sigma)^2)^{N/2}} e^{-\|x\|^2/2(\sqrt{2}\sigma)^2} dx,
\end{aligned}$$

where the last step comes from integrating over all of $\mathbb{R}^N$ and changing variables $y - \mathcal{A}(\theta) \mapsto x$. This last integral calculates the expected squared length of a vector in $\mathbb{R}^N$ with independent $\mathcal{N}(0, 2\sigma^2)$ entries, which is $2N\sigma^2$. Thus, substituting into (53) gives that b has a finite second moment. □

Appendix B.

Here, we compute the phase error vector $\omega := \{\omega_i\}_{i\in V} \subseteq \mathbb{T}$ from a solution to a relaxed version of the feasibility problem (50): for some fixed tolerance $\varepsilon > 0$,

$$\begin{aligned}\text{Find}\quad X \in \mathbb{H}^{n\times n} \qquad \text{such that}\quad & \operatorname{rank}(X) = 1\\ \text{and}\quad & \|\mathcal{P}_{S^\perp}X\|_{\mathrm{HS}}^2 + \sum_{i=0}^{n-1}\bigl|1 - \langle X, \delta_i\delta_i^*\rangle_{\mathrm{HS}}\bigr|^2 \le \varepsilon.\end{aligned}$$

Note that the outer product of any modulation of the optimal ω is necessarily feasible, and so relaxing the problem to make it convex implies that any convex combination of such modulations is feasible for the relaxed version:

Theorem B.1. *Suppose $\omega\colon \mathbb{Z}_n \to \mathbb{T}$ such that $\omega\omega^*$ solves a convex relaxation of the feasibility problem* (50). *Then $\sum_{i=0}^{n-1}\lambda_i E^i\omega\omega^*E^{-i}$ is also a solution for every $\{\lambda_i\}_{i=0}^{n-1} \subseteq \mathbb{R}$ such that $\sum_{i=0}^{n-1}\lambda_i = 1$.*

To be clear, the modulation operator E in Theorem B.1 is the $n \times n$ diagonal matrix whose kth diagonal entry is $e^{2\pi i k/n}$. Before proving the theorem, however, we first consider some intermediate results.

Lemma B.2. *Let $G = (V, E)$ be a circulant graph with n vertices such that n is an odd prime and $\omega\colon V \to \mathbb{T}$. Then $\|\mathcal{P}_{S^\perp}(E\omega\omega^*E^*)\|_{\mathrm{HS}} = \|\mathcal{P}_{S^\perp}(\omega\omega^*)\|_{\mathrm{HS}}$.*

Proof. By Lemma 5.7, G decomposes into copies of the n-cycle. For each resultant n-cycle C, define the matrices X_C and X_{-C} as in (49) and recall the subspace $S = \operatorname{span}\{X_C, X_{-C}\}_{C\subseteq G} + T$, where $T := \{X \in \mathbb{H}^{n\times n} : X[i,j] = 0 \text{ whenever } A[i,j] = 1\}$. Since no two n-cycles in the decomposition of G share an edge and each nonzero entry of X_C and X_{-C} is of unit-modulus, it follows that $\{\frac{1}{\sqrt{n}}X_C, \frac{1}{\sqrt{n}}X_{-C}\}_{C\subseteq G}$ is an orthonormal set. Noting that for $i, j \in V$ we have

$$S = \operatorname{span}\bigl(\{X_C, X_{-C}\}_{C\subseteq G} \cup \{\delta_i\delta_j^*\}_{i\not\leftrightarrow j}\bigr)$$

then implies that the set $\{\frac{1}{\sqrt{n}}X_C, \frac{1}{\sqrt{n}}X_{-C}\}_{C\subseteq G} \cup \{\delta_i\delta_j^*\}_{i\not\leftrightarrow j}$ forms an orthonormal basis for S.

At this point, note that the Pythagorean Theorem gives

$$\begin{aligned} \|\omega\omega^*\|_{\mathrm{HS}}^2 &= \|\mathcal{P}_S(\omega\omega^*)\|_{\mathrm{HS}}^2 + \|\mathcal{P}_{S^\perp}(\omega\omega^*)\|_{\mathrm{HS}}^2, \\ \|E\omega\omega^*E^*\|_{\mathrm{HS}}^2 &= \|\mathcal{P}_S(E\omega\omega^*E^*)\|_{\mathrm{HS}}^2 + \|\mathcal{P}_{S^\perp}(E\omega\omega^*E^*)\|_{\mathrm{HS}}^2, \end{aligned}$$

and so to establish the result it suffices to show the equalities $\|E\omega\omega^*E^*\|_{\mathrm{HS}} = \|\omega\omega^*\|_{\mathrm{HS}}$ and $\|\mathcal{P}_S(E\omega\omega^*E^*)\|_{\mathrm{HS}} = \|\mathcal{P}_S(\omega\omega^*)\|_{\mathrm{HS}}$. To this end, we first obtain

$$\begin{aligned} \|E\omega\omega^*E^*\|_{\mathrm{HS}}^2 &= \mathrm{Tr}[(E\omega\omega^*E^*)^*E\omega\omega^*E^*] = \mathrm{Tr}[E\omega\omega^*E^*E\omega\omega^*E^*] \\ &= \mathrm{Tr}[E^*E\omega\omega^*E^*E\omega\omega^*] = \mathrm{Tr}[(\omega\omega^*)^*\omega\omega^*] = \|\omega\omega^*\|_{\mathrm{HS}}^2, \end{aligned}$$

at which point taking square roots gives the former equality. For the latter, we start by again applying the Pythagorean Theorem using the basis for S given above:

$$\begin{aligned} \|\mathcal{P}_S(E\omega\omega^*E^*)\|_{\mathrm{HS}}^2 = \sum_{C\subseteq G}\Big(|\langle E\omega\omega^*E^*, \tfrac{1}{\sqrt{n}}X_C\rangle_{\mathrm{HS}}|^2 + |\langle E\omega\omega^*E^*, \tfrac{1}{\sqrt{n}}X_{-C}\rangle_{\mathrm{HS}}|^2\Big) \\ + \sum_{i\not\leftrightarrow j}|\langle E\omega\omega^*E^*, \delta_i\delta_j^*\rangle_{\mathrm{HS}}|^2. \qquad (54) \end{aligned}$$

To simplify this expression, notice that

$$\langle E\omega\omega^*E^*, X_C\rangle_{\mathrm{HS}} = \mathrm{Tr}[X_C^*E\omega\omega^*E^*] = \mathrm{Tr}[E^*X_C^*E\omega\omega^*] = \langle \omega\omega^*, E^*X_CE\rangle_{\mathrm{HS}},$$

where $(E^*X_C^*E)[i,j] = e^{2\pi \mathrm{i}(j-i)/n}X_C[i,j]$. A similar computation for X_{-C} yields $(E^*X_{-C}^*E)[i,j] = e^{2\pi\mathrm{i}(i-j)/n}X_{-C}[i,j]$, and so it follows that

$$\langle E\omega\omega^*E^*, X_{\pm C}\rangle_{\mathrm{HS}} = \langle \omega\omega^*, E^*X_{\pm C}E\rangle_{\mathrm{HS}} = e^{\pm 2\pi\mathrm{i}(j-i)/n}\langle\omega\omega^*, X_{\pm C}\rangle_{\mathrm{HS}}. \qquad (55)$$

On the other hand, since $E^*\delta_i = e^{-2\pi\mathrm{i}i/n}\delta_i$, we have

$$\begin{aligned}\langle E\omega\omega^* E^*, \delta_i\delta_j^*\rangle_{\mathrm{HS}} &= \mathrm{Tr}[(\delta_i\delta_j^*)^* E\omega\omega^* E^*] = \mathrm{Tr}[(E^*\delta_i\delta_j^* E)^*\omega\omega^*] \\ &= e^{2\pi\mathrm{i}(i-j)/n}\,\mathrm{Tr}[(\delta_i\delta_j^*)^*\omega\omega^*] = e^{2\pi\mathrm{i}(i-j)/n}\langle\omega\omega^*, \delta_i\delta_j^*\rangle_{\mathrm{HS}} \qquad (56)\end{aligned}$$

for any $i, j \in \{0, \dots, n-1\}$. Substituting (55) and (56) into (54) then yields

$$\begin{aligned}\|\mathcal{P}_S(E\omega\omega^* E^*)\|_{\mathrm{HS}}^2 &= \sum_{C\subseteq G}\Big(|\langle\omega\omega^*, \tfrac{1}{\sqrt{n}}X_C\rangle_{\mathrm{HS}}|^2 + |\langle\omega\omega^*, \tfrac{1}{\sqrt{n}}X_{-C}\rangle_{\mathrm{HS}}|^2\Big) \\ &\quad + \sum_{i\not\leftrightarrow j}|\langle\omega\omega^*, \delta_i\delta_j^*\rangle_{\mathrm{HS}}|^2 \\ &= \|\mathcal{P}_S(\omega\omega^*)\|_{\mathrm{HS}}^2,\end{aligned}$$

which is the desired result. □

Lemma B.3. *Let $G = (V, E)$ be a circulant graph with n vertices such that n is an odd prime and $\omega\colon V \to \mathbb{T}$. For any fixed $\varepsilon > 0$, suppose that*

$$\|\mathcal{P}_{S^\perp}(\omega\omega^*)\|_{\mathrm{HS}}^2 + \sum_{i=0}^{n-1}\big|1 - \langle\omega\omega^*, \delta_i\delta_i^*\rangle_{\mathrm{HS}}\big|^2 \leq \varepsilon.$$

Then

$$\Big\|\mathcal{P}_{S^\perp}\sum_{i=0}^{n-1}\lambda_i E^i\omega\omega^* E^{-i}\Big\|_{\mathrm{HS}}^2 + \sum_{i=0}^{n-1}\Big|1 - \big\langle\sum_{j=0}^{n-1}\lambda_j E^j\omega\omega^* E^{-j}, \delta_i\delta_i^*\big\rangle_{\mathrm{HS}}\Big|^2 \leq \varepsilon$$

for every $\{\lambda_i\}_{i=0}^{n-1} \subseteq \mathbb{R}$ such that $\sum_{i=0}^{n-1}\lambda_i = 1$.

Proof. Suppose $\{\lambda_i\}_{i=0}^{n-1} \subseteq \mathbb{R}$ such that $\sum_{i=0}^{n-1} \lambda_i = 1$ and let $Y := \sum_{i=0}^{n-1} \lambda_i E^i \omega\omega^* E^{-i}$. By the triangle inequality we have

$$\begin{aligned} \|\mathcal{P}_{S^\perp} Y\|_{\text{HS}} &\leq \sum_{i=0}^{n-1} \lambda_i \|\mathcal{P}_{S^\perp}(E^i \omega\omega^* E^{-i})\|_{\text{HS}} \\ &= \|\mathcal{P}_{S^\perp}(E^i \omega\omega^* E^{-i})\|_{\text{HS}} \sum_{i=0}^{n-1} \lambda_i = \|\mathcal{P}_{S^\perp}(E^i \omega\omega^* E^{-i})\|_{\text{HS}}, \end{aligned}$$

and so applying Lemma B.2 yields $\|\mathcal{P}_{S^\perp} Y\|_{\text{HS}} \leq \|\mathcal{P}_{S^\perp}(\omega\omega^*)\|_{\text{HS}}$. Moreover, for any $i \in \{0, \ldots, n-1\}$ we have

$$\begin{aligned} \langle Y, \delta_i \delta_i^* \rangle_{\text{HS}} &= \sum_{j=0}^{n-1} \lambda_j \langle E^j \omega\omega^* E^{-j}, \delta_i \delta_i^* \rangle_{\text{HS}} \\ &= \sum_{j=0}^{n-1} \lambda_j \langle \omega\omega^*, E^{-j} \delta_i \delta_i^* E^j \rangle_{\text{HS}} \\ &= \langle \omega\omega^*, \delta_i \delta_i^* \rangle_{\text{HS}} \sum_{j=0}^{n-1} \lambda_j = \langle \omega\omega^*, \delta_i \delta_i^* \rangle_{\text{HS}}, \end{aligned}$$

and so it follows that

$$\begin{aligned} &\|\mathcal{P}_{S^\perp} Y\|_{\text{HS}}^2 + \sum_{i=0}^{n-1} \big|1 - \langle Y, \delta_i \delta_i^* \rangle_{\text{HS}}\big|^2 \\ &\qquad \leq \|\mathcal{P}_{S^\perp}(\omega\omega^*)\|_{\text{HS}}^2 + \sum_{i=0}^{n-1} \big|1 - \langle \omega\omega^*, \delta_i \delta_i^* \rangle_{\text{HS}}\big|^2 \leq \varepsilon, \end{aligned}$$

completing the proof. □

Proof of Theorem B.1. Suppose $\{\lambda_i\}_{i=0}^{n-1} \subseteq \mathbb{R}$ such that $\sum_{i=0}^{n-1} \lambda_i = 1$ and let $Y := \sum_{i=0}^{n-1} \lambda_i E^i \omega\omega^* E^{-i}$. Since Y is a convex combination of solutions to the feasibility problem (50), it is feasible for any convex relaxation provided the norm and diagonal constraints are satisfied. To this end, note that Lemma B.3 implies that $\|\mathcal{P}_{S^\perp} Y\|_{\text{HS}}^2 + \sum_{i=0}^{n-1} \big|1 - \langle Y, \delta_i \delta_i^* \rangle_{\text{HS}}\big|^2 \leq \varepsilon$, and so Y is feasible. □

To reiterate, it is not clear how to obtain a solution to the feasibility problem (50); indeed, properly relaxing it to a convex feasibility problem is still an open problem. Assuming such a relaxation exists, recovering the desired phase vector from a solution to the relaxed version is possible. To see this, let ω be a given vector of unknown phase errors and suppose the matrix Y solves a convex relaxation of (50) (without noise) for some $\varepsilon > 0$. Then, as Theorem B.1 indicates, we know that $Y = \sum_{i=0}^{n-1} \lambda_i E^i \omega\omega^* E^{-i}$ for some $\{\lambda_i\}_{i=0}^{n-1} \subseteq \mathbb{R}$ such that $\sum_{i=0}^{n-1} \lambda_i = 1$. In order to recover the actual phase vector ω, we therefore need some way of decomposing the matrix Y. Notice that if the vectors $\{E^i\omega\}_{i=0}^{n-1}$ are mutually orthogonal, then Y is diagonalizable by the spectral theorem. Such a diagonalization would have the effect of arranging normalized versions of $\{E^i\omega\}_{i=0}^{n-1}$ as the columns of some matrix, which we could then analyze. For this approach, we require the following lemma:

Lemma B.4. *Let $\omega\colon \mathbb{Z}_n \to \mathbb{T}$ be a vector of unit-modulus phases. Then the set of modulations $\{\frac{1}{\sqrt{n}} E^i\omega\}_{i=0}^{n-1}$ is orthonormal.*

Proof. For any $i, j \in \{0, \dots, n-1\}$, consider the inner-product

$$\langle E^i\omega, E^j\omega\rangle_{\mathrm{HS}} = \langle \omega, E^{j-i}\omega\rangle_{\mathrm{HS}} = \mathrm{Tr}[(E^{j-i}\omega)^*\omega] = \omega^* E^{i-j}\omega = \sum_{k=0}^{n-1} e^{2\pi \mathrm{i}(j-i)k/n}|\omega_k|^2.$$

Since $|\omega_k| = 1$ for all $k \in \{0, \dots, n-1\}$, the geometric sum formula then yields

$$\langle E^i\omega, E^j\omega\rangle_{\mathrm{HS}} = \begin{cases} n & \text{if } i = j \\ \frac{e^{2\pi \mathrm{i}(j-i)}-1}{e^{2\pi \mathrm{i}(j-i)/n}-1} & \text{if } i \neq j, \end{cases}$$

from which it follows that

$$\langle \tfrac{1}{\sqrt{n}} E^i\omega, \tfrac{1}{\sqrt{n}} E^j\omega\rangle_{\mathrm{HS}} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \qquad \square$$

For the matrix Y above, the result of Lemma B.4 implies the spectral decomposition

$$Y = \sum_{i=0}^{n-1} \lambda_i (E^i\omega)(E^i\omega)^* = V(n\Lambda)V^*,$$

where V is the matrix whose ith column is the vector $\frac{1}{\sqrt{n}}E^i\omega$ for each $i = 0, \ldots, n-1$ and Λ is the diagonal matrix whose entries are $\Lambda[i,i] = \lambda_i$. Note that we may arbitrarily permute the columns of V without changing the matrix Y. Similarly, multiplying each column by an arbitrary phase in $\mathbb{T}$ has no effect on Y, and so the most general expression for a solution to (50) is of the form $Y = (VP\Psi)(n\Lambda)(VP\Psi)^*$ for some permutation matrix P and diagonal matrix of phases Ψ.

Now consider the diagonal matrix W defined by $W[i,i] = \omega_i$ for each $i = 1, \ldots, n-1$. If w denotes the n-vector whose coordinates are all 1, then we may decompose V in terms of W and the n-dimensional inverse DFT matrix F:

$$V = \tfrac{1}{\sqrt{n}}W \left[\begin{array}{cccc} E^0w & E^1w & \cdots & E^{n-1}w \end{array} \right] = WF.$$

Hence, columns of the matrix $U := WFP\Psi$ are each of the form $\{\frac{1}{\sqrt{n}}e^{2\pi \mathrm{i} z}e^{2\pi \mathrm{i} k/n}\omega_k\}_{k=0}^{n-1}$ for some $z \in \mathbb{R}/\mathbb{Z}$, i.e., the columns of U are modulated versions of ω, multiplied by distinct global phases. Since modulation and global phase are the two degrees of freedom associated with recovering the vector ω (see Lemma 5.8), it follows that any column of U (properly scaled) yields an acceptable estimate for ω.

In the presence of noise, we desire a more stable solution than simply pulling any column from U. Indeed, since any column of U is acceptable, it makes sense that each column would be equally perturbed by noise. To leverage this, we may attempt to "integrate" over all columns in order to average out the effect of noise. This is a common approach to stable algorithms and is motivated by the Central Limit Theorem. One such way of doing this is summarized in Algorithm 6.

```
Algorithm 6
Input: Matrix solution $Y = U(n\Lambda)U^*$ to a relaxed version of problem (50)
Output: Vector $\omega$ of phases such that $Y = \sum_{i=0}^{n-1} \lambda_i E^i \omega\omega^* E^{-i}$ for some convex sum $\sum_{i=0}^{n-1} \lambda_i = 1$
  Fix an initial estimate $\hat{\omega} \leftarrow \sqrt{n} u_0$, where $u_0$ is the first column of $U$
  Fix a threshold $\varepsilon > 0$
  Initialize an $n \times n$ matrix $P$ of zeros
  for $i, j = 0$ to $n - 1$ do
    if $|1 - \frac{1}{\sqrt{n}} |\langle E^i \omega, u_j \rangle|| \leq \varepsilon$ then
      $P[i, j] \leftarrow 1$          {Detect the permutation matrix which orders the columns of $U$ as consecutive modulations of $\hat{\omega}$}
    end if
  end for
  Compute $A \leftarrow \sqrt{n}(UP^{-1}) \circ F^*$          {$F^*$ denotes the $n \times n$ DFT matrix; $\circ$ is the Hadamard matrix product}
  Compute the singular value decomposition $Q\Sigma R^* \leftarrow \mathrm{SVD}[A]$
  Compute the updated estimate $\hat{\omega} \leftarrow q_0$, where $q_0$ is the first column of $Q$
  Output: $\omega = \{\hat{\omega}_i / |\hat{\omega}_i|\}_{i=0}^{n-1}$
```

This algorithm is quite natural considering the above discussion in which we decomposed the matrix Y. Indeed, since $U = VP\Psi$ for some permutation matrix P and diagonal matrix Ψ of phases, where $V = WF$ is in terms of the $n \times n$ inverse DFT, the matrix A in Algorithm 6 approximates the outer product $\omega\psi^T$, where ψ is the diagonal of Ψ. Thus, the leading column of Q in the singular value decomposition is the desired vector of phase errors from the best low rank approximation for A.

Appendix C.

In the spirit of reproducible research [45], we provide here the code used to run simulations in Chapter V. All data was acquired using versions of the code below.

```
%% Iterative Angular Synchronization
% Determines a phase error vector (up to global phase and modulation)
%    from nested relative phase information. The first level of
%    relative phase is encoded as edge measurements in a graph G; the
%    second level is encoded as the edge measurements in a related graph
%    G^prime. The vertex measurements in G^prime are determined, which
%    then become the edge measurements in G. The resultant vertex
%    measurements in G are the entries of the phase error vector.
% Based on angular synchronization, Amit Singer (2011).



n = 101;            % Specify number of vertices in G (must be prime >2;
                    %    this ensures G decomposes into n-cycles)
c = 10;             % Specify number of n-cycles in the decomposition of
                    %    G (must be <n/2 so that G can be directed)
noise = 0.1;        % Specify magnitude of adversarial noise (must be in
                    %    [0,1]); 0 is no noise, 1 is full N(0,1) noise

firstcolumndone = 0;
while firstcolumndone == 0
    r = zeros(1,n);               % Generates a random vector r of 0s
    indices = randperm(n-1);      %    and 1s which will generate the
    for k = 2 : n                 %    adjacency matrix of G; r will
        if indices(k-1) <= c      %    have c nonzero entries, each
            r(k) = 1;             %    corresponding to a distinct cycle
        end                       %    in G
    end
    revr = zeros(n,1);
    revr(1) = r(1);
    for k = 2 : n
        revr(k) = r(n-k+2);
    end
    if r*revr == 0                % Ensures r will generate a directed
        firstcolumndone = 1;      %    graph G
    end
end
```

```
I = eye(n);
P = zeros(n);
A = zeros(n);

for k = 1 : n                          % Builds the permutation matrix
    if k < n                           %    used to generate the
        P(:,k) = P(:,k) + I(:,k+1);    %    adjacency matrix of G
    else
        P(:,k) = P(:,k) + I(:,1);
    end
end

for k = 1 : n                        % Generates the adjacency matrix
    A(:,k) = A(:,k) + P^(k-1)*r';    %    A of G; A is circulant by
end                                  %    construction

Ac = zeros(n,n,c);

count = 0;
rc = zeros(1,n);
for k = 2 : n                 % Generates the adjacency matrices of each
    if r(k) == 1              %    cycle in G; the kth cycle adjacency
        count = count + 1;    %    matrix is stored as Ac(:,:,k)
        rc(k) = 1;
        for j = 1 : n
            Ac(:,j,count) = Ac(:,j,count) + P^(j-1)*rc';
        end
        rc = 0*rc;
    end
end

r1 = rand(n,1);
omega1 = exp(2*pi*1i*r1);     % Generates a random phase error vector;
                              %    this is the true vertex data for G
exact = omega1*omega1';

Xc = zeros(n,n,c);

for k = 1 : c                        % Generates the true weighted
    Xc(:,:,k) = Ac(:,:,k).*exact;    %    adjacency matrices for each
end                                  %    cycle in G
```

```
sigma = zeros(n,c);                                                         81

for k = 1 : c                              % Lines 85-93 pull the nonzero   83
    for j = 1 : n                          %    entries from each Xc and
        for m = 1 : n                      %    organize them in vector      85
            if Xc(j,m,k) ~= 0              %    form; for the kth cycle,
                sigma(j,k) = Xc(j,m,k);    %    sigma(:,k) is the nonzero    87
            end                            %    data from Xc(:,:,k)
        end                                                                  89
    end
end                                                                          91

Sigma = zeros(n,c);                                                          93

for k = 1 : c                    % Reorders the nonzero data from each       95
    shift = 0;                   %    Xc so that consecutive entries
    x = 0;                       %    come from consecutive edges in         97
    while shift == 0             %    each cycle of G^prime; the kth
        x = x+1;                 %    column of Sigma represents the         99
        if Ac(1,x,k) == 1;       %    true edge measurements for the kth
            shift = x-1;         %    cycle; this data is the received       101
        end                      %    relative phases uncorrupted by
    end                          %    noise                                  103
    for j = 1 : n
        Sigma(j,k) = sigma(1+mod((j-1)*shift,n),k);                          105
    end
end                                                                          107

Sigma1 = zeros(n,c);                                                         109

for k = 1 : c                                                                111
    Sigma1(:,k) = Sigma(:,k) + noise*(randn(n,1)+1i*randn(n,1));
    Sigma1(:,k) = Sigma1(:,k)./abs(Sigma1(:,k));                             113
end                % Adds N(0,1) noise to the input signal according to
                   %    the predefined parameter noise                       115

data = zeros(n,n,c);                                                         117

for k = 1 : c         % Creates the weighted adjacency matrices for each     119
    for j = 1 : n     %    cycle in G^prime; nested relative phases for the
        if j < n      %    kth cycle are stored in data(:,:,k)               121
            data(j,j+1,k) = Sigma1(j,k)*conj(Sigma1(j+1,k));
        else                                                                 123
```

```
            data(j,1,k) = Sigma1(j,k)*conj(Sigma1(1,k));
        end
    end
end


%
% From this point forward the algorithm is reconstructing the phase error
%    vector from noise-corrupted, nested relative phase information; in
%    practice, this is where real data will enter the algorithm
%


sigma1 = zeros(n,c);

for k = 1 : c                                % Angular synchronization; the
    [V,~] = eig(data(:,:,k)+data(:,:,k)');   %    columns of sigma1 are
    sigma1(:,k) = V(:,n)./abs(V(:,n));       %    estimates of corresponding
end                                          %    columns of Sigma (up to
                                             %    distinct global phases)

% To check accuracy, note that
%         (sigma1(:,k)*sigma1(:,k)')-(Sigma(:,k)*Sigma(:,k)')
%    should be a matrix of zeros for each k (without noise)

sigma2 = zeros(n,c);

for k = 1 : c                        % Reorders the input data such that
    shift = 0;                       %    consecutive entries correspond to
    x = 0;                           %    consecutive rows in the weighted
    while shift == 0                 %    adjacency matrices of each cycle
        x = x+1;                     %   G^prime; note this is the inverse
        if Ac(1,x,k) == 1;           %    process to that performed above
            shift = x-1;
        end
    end
    for j = 1 : n
        sigma2(1+mod((j-1)*shift,n),k) = sigma1(j,k);
    end
end


estimate = zeros(n,n,c);
```

```

for k = 1 : c
    estimate(:,:,k) = diag(sigma2(:,k))*Ac(:,:,k);
end                 % Generates weighted adjacency matricesfor each cycle
                    %    in G; the kth cycle is stored as estimate(:,:,k)

s = zeros(c);

count1 = 0;
count2 = 0;

for k = 2 : n                             % Generates a matrix s of exponents;
    if r(k) == 1                          %    exponent s(k,j) represents the
        count1 = count1 + 1;              %    power such that estimate(:,:,k)
        for j = 2 : n                     %    and estimate(:,:,j)^s(k,j)
            if r(j) == 1                  %    have common support; note that
                count2 = count2 + 1;      %    s(k,j)s(j,k)=1(mod n)
                for m = 1 : n
                    if mod((k-1)*(m-1),n) == 1
                        inverse = m-1;
                    end
                end
                s(count2,count1) = mod((j-1)*inverse,n);
            end
        end
        count2 = 0;
    end
end

% Below (commented out) is an alternative way of computing the matrix s

% s1 = zeros(c);
%
% for k = 1 : c
%     for j = 1 : c
%         for power = 1 : n
%             if Ac(:,:,j)^power == Ac(:,:,k)
%                 s1(k,j) = power;
%             end
%         end
%     end
% end

```

```
alpha = zeros(c);

for k = 1 : c
    for j = 1 : c
        sumsum = sum(sum(estimate(:,:,j)^(s(k,j)).*conj(estimate(:,:,k))));
        alpha(k,j) = sumsum/abs(sumsum);
    end            % Computes phases such that
end                %    alpha(k,j)estimate(:,:,k) = estimate(:,:,j)^s(k,j)
                   %    for all k,j=1,...,c

% To check accuracy, note that
%       alpha(k,j)*estimate(:,:,k)-estimate(:,:,j)^(s(k,j))
%    should be a matrix of zeros for each pair (k,j) (without noise)


%
% At this point, note that we have the edge measurements for each cycle in
%    G up to distinct global phase factors; to reconcile these global
%    phases, observe that in the noiseless case we have
%        alpha(k,j)estimate(:,:,k) = estimate(:,:,j)^s(k,j)
%    and so, letting Estimate(:,:,k) denote the actual weighted adjacency
%    matrices, there exist phase factors beta(k) such that
%        alpha(k,j)Estimate(:,:,k)/beta(k)
%           = (Estimate(:,:,j)/beta(j))^s(k,j)
%    from which equality in the noiseless case implies that
%        alpha(k,j)/beta(k) = 1/beta(j)^s(k,j)
%    we then immediately have the relation
%         beta(k) = alpha(k,1)beta(1)^s(k,1)
%


beta = zeros(c,1);

matrix = (estimate(:,:,1) == 0) + estimate(:,:,1);
number = prod(prod(matrix));
beta(1) = number^(-1/n);      % Fixes beta(1) to be the inverse of the
                              %    product of the phases in estimate(:,:,1)

for k = 2 : c
    beta(k) = alpha(k,1)*beta(1)^(s(k,1));
end                 % Populates the phase vector beta based on alphas
                    %    and relationship above using beta(1)
```

```
XcEst = zeros(n,n,c);

for k = 1 : c
    XcEst(:,:,k) = estimate(:,:,k)*(conj(beta(k)));
end        % Generates the weighted adjacency matrices of the cyles in G
           %    such that edge measurements in all cycles are estimated up
           %    to a common global phase factor

A1 = zeros(n);

for k = 1 : c                        % Generates the weighted adjacency matrix
    A1 = A1 + XcEst(:,:,k);          %    A1 of G (up to a global phase)
end

lead = 1;
leader = 0;
for k = 1 : n                                   % Eigenvector method for
    psi = exp(2*pi*1i*k/n);                     %    angular synchronization
    H = psi*A1;                                 %    peformed for each phase
    [V,~] = eig(H+H');                          %    factor psi
    v = abs(V(:,n)) - sqrt(1/n)*ones(n,1);
    discrepancy = max(abs(v));                  % Determines the phase psi
    if discrepancy < lead                       %    such that the leading
        lead = discrepancy;                     %    eigenvector of H+H' is
        leader = k;                             %    closest to unimodular
    end
end

psi = exp(2*pi*1i*leader/n);          % Angular synchronization using the
H = psi*A1;                           %    optimal phase psi generated above
[V,~] = eig(H+H');

omega = V(:,n)./abs(V(:,n));          % Generates the estimated phase error
                                      %    vector omega


% Last edited: 17 Feb 2014
% Edited by: Aaron A. Nelson
```

```
%% Modulation and Global Phase Detection
%  Determines the best moduation and phase factor for reconstructing an
%     image from multistatic SAR data with phase errors known up to a
%     global phase and modulation.


%
% Before running this code, load a matrix X of a SAR image.
%


X = im2double(X(:,:,1));
[width height] = size(X);

n = 101;                  % Specify size of partition of the Fourier
                          %    domain; this should correspond to the
                          %    number measurements taken (i.e., the
                          %    length of the phase error vector) and
                          %    must be prime >2 to correspond to usable
                          %    circulant graphs
noise = 0.1;              % Specify magnitude of noise (must be in [0,1]);
                          %    0 is no noise, 1 is full complex N(0,1) noise

mask = zeros(width,height);

for k = 1 : width               % Partitions Fourier domain into n sections
    if k <= width/2             %    by angle; each section corresponds to
        x = k-1/2;              %    a slice of the Fourier transform data
    else                        %    that is available from multistatic SAR
        x = k-width-1/2;
    end
    for j = 1 : height
        if j <= height/2
            y = j-1/2;
        else
            y = j-height-1/2;
        end
        angle = atan(y/x);
        mask(k,j) = floor((angle+pi/2)*n/pi) + 1;
    end
end

```

```
I = eye(n);

randphases = ones(n,1) + noise*(randn(n,1)+1i*randn(n,1));
randphases = randphases./abs(randphases);
     % Generates random phases close to 1 with noise proportional to
     %    predefined parameter noise (complex N(0,1), centered at 1)

globalphase = randn(1)+1i*randn(1);          % Generates a random global
globalphase = globalphase/abs(globalphase);  %    phase

truemodulation = ceil(rand*n);       % Generates a random modulation index

randphases = globalphase*randphases.*fft(I(:,truemodulation));
            % Simulates the (noisy) phase error vector, known up to a
            %    global phase and modulation

X = circshift(X,[floor(width/2),floor(height/2)]);
Y = fft2(X);                         % Centers the image around the "origin"
                                     %    and takes the 2D Fourier transform
Ynoisy = Y.*randphases(mask);
        % Applies noisy phases to each slice of the image in the Fourier
        %    domain; this simulates the input data from multistatic SAR


%
% From this point forward the algorithm is recovering the modulation and
%    global phase factor from simulated multistatic SAR data constructed
%    above; this is where real data will enter the algorithm
%


tvnorms = zeros(n,1);

for modulation = 1 : n                       % Determines an estimate for
    phases = conj(fft(I(:,modulation)));     %    the modulation by
    Ymodulated = Ynoisy.*phases(mask);       %    maximizing total variation
    Xhat = ifft2(Ymodulated);                %    in the image over all
    Xhat = abs(Xhat);                        %    possible modulations
    tvnorms(modulation)...
        = sum(sum(abs(Xhat-circshift(Xhat,[1,0]))))...
          + sum(sum(abs(Xhat-circshift(Xhat,[0,1]))));
end             % Note this implicitly ignores the global phase factor
```

```
[value modulationestimate] = max(tvnorms);
    % Determines the modulation index that maximizes the total variation

phases = conj(fft(I(:,modulationestimate)));  % Computes the estimated
Ymodulated = Ynoisy.*phases(mask);             %    image with the estimate
Xhat = ifft2(Ymodulated);                      %    for the modulation

globalphaseestimate = sum(sum(Xhat));
globalphaseestimate = globalphaseestimate/abs(globalphaseestimate);
        % Determines an estimate for the global phase factor; takes the
        %    estimate to be the average phase of all data in the
        %    estimated image

Xhat = real(Xhat*conj(globalphaseestimate));
        % The final estimate for the image, with the estimated modulation
        %    and global phase

globalphaseerror = abs(globalphaseestimate-globalphase);
relativeerror = norm(Xhat-X)/norm(X);


% Last edited: 18 Feb 2014
% Edited by: Aaron A. Nelson
```

Bibliography

1. M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions, Dover Publications, New York, 1964.

2. E. Aguilera, S. Baumgartner, I. Hajnsek, R. Horn, T. Jagdhuber, M. Jäger, A. Moreira, M. Nannini, A. Nottensteiner, K. Papathanassiou, P. Prats, A. Reigber, R. Scheiber, Very-high resolution airborne synthetic aperture radar imaging: Signal processing and applications, Proc. IEEE 101 (2013) 759–783.

3. B. Alexeev, A. S. Bandeira, M. Fickus, D. G. Mixon, Phase retrieval with polarization, SIAM J. Imaging Sci. 7 (2014) 35–66.

4. B. Alexeev, J. Cahill, D. G. Mixon, Full spark frames, J. Fourier Anal. Appl. 18 (2012) 1167–1194.

5. R. Balan, Reconstruction of signals from magnitudes of redundant representations, Available online: 1207.1134

6. R. Balan, B. G. Bodmann, P. G. Casazza, D. Edidin, Fast algorithms for signal reconstruction without phase, Proc. SPIE 6701, Wavelets XII (2007) 67011L.

7. R. Balan, B. G. Bodmann, P. G. Casazza, D. Edidin, Painless reconstruction from magnitudes of frame coefficients, J. Fourier Anal. Appl. 15 (2009) 488–501.

8. R. Balan, P. Casazza, D. Edidin, On signal reconstruction without phase, Appl. Comp. Harmon. Anal. 20 (2006) 345–356.

9. A. S. Bandeira, Y. Chen, D. G. Mixon, Phase retrieval from power spectra of masked signals, Available online: arXiv:1303.4458

10. A. S. Bandeira, J. Cahill, D. G. Mixon, A. A. Nelson, Fundamental limits of phase retrieval, In: Proc. 10th Int. Conf. Sampling Theory Appl., Bremen, Germany, (2013) 77–80.

11. A. S. Bandeira, J. Cahill, D. G. Mixon, A. A. Nelson, Saving phase: Injectivity and stability for phase retrieval, to appear in Appl. Comput. Harmon. Anal.

12. R. Basri, O. Ozyesil, A. Singer, Camera motion estimation by convex programming, Available online: arXiv:1312.5047

13. R. Beard, D. W. Casbeer, A. L. Swindlehurst, Connectivity in a UAV multistatic radar network, In: AIAA Guidance Navig. Control Conf., Keystone, CO (2006).

14. J. J. Benedetto, M. Fickus, Finite normalized tight frames, Adv. Comput. Math. 18 (2003) 357–385.

15. P. Bezoušek, V. Schejbal, Bistatic and multistatic radar systems, Radioengineering, 17 (2008) 53–59.

16. B. G. Bodmann, P. G. Casazza, The road to equal-norm Parseval frames, J. Funct. Anal. 258 (2010) 397–420.

17. B. G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames, Available online: arXiv:1302.5487

18. W. M. Brown, SAR resolution in the presence of phase errors, IEEE Trans. Aerosp. Electron. Syst. 24 (1988) 808–814.

19. O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D. K. Satapathy, J. F. van der Veen, Diffractive imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels, Acta Cryst. A63 (2007) 306–314.

20. J. Cahill, M. Fickus, D. G. Mixon, M. J. Poteet, N. Strawn, Constructing finite frames of a given spectrum and set of lengths, Appl. Comput. Harmon. Anal. 35 (2013) 52–73.

21. T. M. Calloway, G. W. Donohoe, Subaperture autofocus for synthetic aperture radar, IEEE Trans. Aerosp. Electron. Syst. 30 (1994) 617–621.

22. Ç. Candan, M. A. Kutay, H. M. Ozaktas, The discrete fractional Fourier transform, IEEE Trans. Signal Process. 48 (2000) 1329–1337.

23. E. J. Candès, The restricted isometry property and its implications for compressed sensing, C. R. Acad. Sci. Paris, Ser. I 346 (2008) 589–592.

24. E. J. Candès, Y. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion, SIAM J. Imaging Sci. 6 (2013) 199–225.

25. E. J. Candès, X. Li, Solving quadratic equations via PhaseLift when there are about as many equations as unknowns, Available online: arXiv:1208.6247

26. E. J. Candès, X. Li, M. Soltanolkotabi, Phase retrieval from coded diffraction patterns, Available online: arXiv:1310.3240

27. E. J. Candès, T. Strohmer, V. Voroninski, PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming, Commun. Pure Appl. Math. 66 (2013) 1241–1274.

28. W. G. Carrara, R. S. Goodman, R. M. Majewski, Spotlight synthetic aperture radar–signal processing algorithms, Artech House, Norwood, MA (1995).

29. P. G. Casazza, M. Fickus, D. G. Mixon, Auto-tuning unit norm frames, Appl. Comput. Harmon. Anal. 32 (2012) 1–15.

30. P. G. Casazza, M. Fickus, D. G. Mixon, Y. Wang, Z. Zhou, Constructing tight fusion frames, Appl. Comput. Harmon. Anal. 30 (2011) 175–187.

31. P. G. Casazza, M. Fickus, J. C. Tremain, E. Weber, The Kadison-Singer problem in mathematics and engineering: A detailed account, Contemporary Math., 414

Operator theory, operator algebras and applications, D. Han, P.E.T. Jorgensen and D.R. Larson (Eds.) (2006) 297–356.

32. P. G. Casazza, J. Kovačević, Equal-norm tight frames with erasures, Adv. Comput. Math. 18 (2003) 387–430.

33. D. W. Casbeer, A. L. Swindlehurst, P. Zhan, A centralized control algorithm for target tracking with UAVs, In: Proc. 39th IEEE Asilomar Conf. Signals Syst. Comput., Pacific Grove, CA (2005) 1148–1152.

34. A. Chai, M. Moscoso, G. Papanicolaou, Array imaging using intensity-only measurements, Inverse Probl. 27 (2011) 015005.

35. Z. Chen, J. J. Dongarra, Condition numbers of Gaussian random matrices, SIAM J. Matrix Anal. Appl. 27 (2005) 603–620.

36. M. Cheney, B. Yazici, Radar imaging with independently moving transmitters and receivers, In: Defense Adv. Signal Process., (2006).

37. H. T. Chuah, V. C. Koo, T. S. Lim, A comparison of autofocus algorithms for SAR imagery, Progr. Electromagn. Res. Symp. 1 (2005) 16–19.

38. A. Conca, D. Edidin, M. Hering, C. Vinzant, An algebraic characterization of injectivity in phase retrieval, Available online: arXiv:1312.0158

39. S. Cook, The P versus NP problem, Available online: `http://www.claymath.org/millennium/PvsNP/pvsnp.pdf`

40. J. C. Dainty, J. R. Fienup, Phase retrieval and image reconstruction for astronomy, In: H. Stark, ed., Image Recovery: Theory and Application, Academic Press, New York (1987).

41. I. Daubechies, A. Grossmann, Y. Meyer, Painless nonorthogonal expansions, J. Math. Phys. 27 (1986) 1271–1283.

42. K. R. Davidson, S. J. Szarek, Local operator theory, random matrices and Banach spaces, In: W. B. Johnson, J. Lindenstrauss (Eds.), Handbook in Banach Spaces Vol I, Elsevier (2001) 317–366.

43. L. Demanet, P. Hand, Stable optimizationless recovery from phaseless linear measurements, Available online: arXiv:1208.1803

44. M. N. Do, R. L. Morrison, Jr., D. C. Munson, Jr., MCA: A multichannel approach to SAR autofocus, IEEE Trans. Image Process. 18 (2009) 840–853.

45. D. L. Donoho, A. Maleki, I. Rahman, M. Shahram, V. Stodden, Reproducible research in computational harmonic analysis, Computing in Science and Engineering, 11 (2009) 8–18.

46. R. J. Duffin, A. C. Schaeffer, A class of nonharmonic Fourier series, Trans. Amer. Math. Soc. 72 (1952) 341–366.

47. K. Dykema, N. Strawn, Manifold structure of spaces of spherical tight frames, Int. J. Pure Appl. Math. 28 (2006) 217–256.

48. P. H. Eichel, D. C. Ghiglia, C. V. Jakowatz, Jr., P. A. Thompson, D. E. Wahl, Spotlight-mode synthetic aperture radar: A signal processing approach, Kluwer Academic Publishers, Norwell, MA (1996).

49. Y. C. Eldar, S. Mendelson, Phase retrieval: Stability and recovery guarantees, Available online: arXiv:1211.0872

50. A. Fannjiang, Absolute uniqueness in phase retrieval with random illumination, Inverse Probl. 28 (2012) 075008.

51. M. Fickus, D. G. Mixon, Numerically erasure-robust frames, Linear Algebra Appl. 437 (2012) 1394–1407.

52. M. Fickus, D. G. Mixon, A. A. Nelson, Y. Wang, Phase retrieval from very few measurements, to appear in Linear Algebra Appl.

53. M. Fickus, D. G. Mixon, M. J. Poteet, N. Strawn, Constructing all self-adjoint matrices with prescribed spectrum and diagonal, Adv. Comput. Math. 39 (2013) 585-609.

54. J. R. Fienup, J. J. Miller, Aberration correction by maximizing generalized sharpness metrics, J. Opt. Soc. Amer. A 20 (2003) 609–620.

55. J. R. Fienup, J. C. Marron, T. J. Schulz, J. H. Seldin, Hubble Space Telescope characterized by using phase-retrieval algorithms, Appl. Optics 32 (1993) 1747–1767.

56. J. Finkelstein, Pure-state informationally complete and "really" complete measurements, Phys. Rev. A 70 (2004) 052107.

57. S. T. Flammia, A. Silberfarb, C. M. Caves, Minimal informationally complete measurements for pure states, Found. Phys. 35 (2005) 1985–2006.

58. V. K. Goyal, M. Vetterli, N. T. Thao, Quantized overcomplete expansions in $\mathbb{R}^N$: Analysis, synthesis, and algorithms, IEEE Trans. Inform. Theory 44 (1998) 1–31.

59. C. Hanle, E. Pell, Bistatic and multistatic radar, IEE Proc., 133 (1986) 585–586.

60. R. W. Harrison, Phase problem in crystallography, J. Opt. Soc. Am. A 10 (1993) 1046–1055.

61. R. Hartshorne, Algebraic Geometry, Graduate Texts in Mathematics, Springer, New York (1977).

62. T. Heinosaari, L. Mazzarella, M. M. Wolf, Quantum tomography under prior information, Commun. Math. Phys. 318 (2013) 355–374.

63. W. Hong, C. Jiang, Y. Wu, B. Zhang, Z. Zhang, Y. Zhao, Autofocus of sparse microwave imaging radar based on phase recovery, IEEE Int. Conf. Signal Process. Commun. Comput. (2013).

64. P. E. Howland, Target tracking using television-based bistatic radar, IEEE Proc. Radar Sonar Navig., 146 (1999) 166–174.

65. K. Jaganathan, S. Oymak, B. Hassibi, Recovery of sparse 1-D signals from the magnitudes of their Fourier transform, Available online: arXiv:1206.1405

66. I. M. James, Euclidean models of projective spaces, Bull. London Math. Soc. 3 (1971) 257–276.

67. P. Jaming, Uniqueness results for the phase retrieval problem of fractional Fourier transforms of variable order, Available online: arXiv:1009.3418

68. R. M. Karp, Reducibility Among Combinatorial Problems, In: R. E. Miller, J. W. Thatcher (Eds.), Complexity of Computer Computations (1972) 85–103.

69. S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.

70. L. Khachiyan, On the complexity of approximating extremal determinants in matrices, J. Complexity 11 (1995) 138–153.

71. V. Krishnan, J. Swoboda, C. E. Yarman, B. Yazici, Multistatic synthetic aperture radar image formation, IEEE Trans. Image Process., (2010) 1290–1306.

72. E. L. Lehmann, G. Casella, Theory of Point Estimation, 2nd ed., Springer, 1998.

73. R. J. Lipton, Definitions, Definitions, Do We Need Them? Godel's Lost Letter and P=NP, Available online: `http://rjlipton.wordpress.com/2010/01/23/definitions-definitions-do-we-need-them/`

74. M. R. Lopez, Synthetic aperture radar applications, Available online: `http://www.sandia.gov/radar/sarapps.html`

75. K. H. Mayer, Elliptische Differentialoperatoren und Ganzzahligkeitssätze für charakteristische Zahlen, Topology 4 (1965) 295–313.

76. J. Miao, T. Ishikawa, Q. Shen, T. Earnest, Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes, Annu. Rev. Phys. Chem. 59 (2008) 387–410.

77. R. J. Milgram, Immersing projective spaces, Ann. Math. 85 (1967) 473–482.

78. R. P. Millane, Phase retrieval in crystallography and optics, J. Opt. Soc. Am. A 7 (1990) 394–411.

79. R. L. Morrison, Jr., D. C. Munson, Jr., An experimental study of a new entropy-based SAR autofocus technique, In: Proc. IEEE Int. Conf. Image Process., Rochester, NY (2008) 441–444.

80. A. Mukherjee, Embedding complex projective spaces in Euclidean space, Bull. London Math. Soc. 13 (1981) 323–324.

81. M Püschel, J. Kovačević, Real, tight frames with maximal robustness to erasures, Proc. Data Compression Conf. (2005) 63–72.

82. H. Sahinoglou, S. D. Cabrera, On phase retrieval of finite-length sequences using the initial time sample, IEEE Trans. Circuits Syst. 38 (1991) 954–958.

83. Sandia National Laboratories, Ku-band synthetic aperture radar imagery, Available online: `http://www.sandia.gov/radar/imageryku.html`

84. A. Singer, Angular synchronization by eigenvectors and semidefinite programming, Appl. Comput. Harmon. Anal. 30 (2011) 20–36.

85. D. L. Sun, J. O. Smith III, Estimating a signal from a magnitude spectrogram via convex optimization, Available online: arXiv:1209.2076

86. B. Steer, On the embedding of projective spaces in euclidean space, Proc. London Math. Soc. 21 (1970) 489–501.

87. A. Vogt, Position and momentum distributions do not determine the quantum mechanical state, In: A. R. Marlow (Ed.), Mathematical Foundations of Quantum Theory, Academic Press, New York (1978).

88. V. Voroninski, A comparison between the PhaseLift and PhaseCut algorithms, Available online: `http://math.berkeley.edu/~vladv/PhaseCutProofs.pdf`

89. V. Voroninski, Phase retrieval from quadratic unitary measurements and implications for Wright's conjecture, Available online: `http://math.berkeley.edu/~vladv/UnitaryCase.pdf`

90. I. Waldspurger, A. d'Aspremont, S. Mallat, Phase recovery, MaxCut and complex semidefinite programming, Available online: arXiv:1206.0102

91. A. Walther, The question of phase retrieval in optics, Opt. Acta 10 (1963) 41–49.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704–0188*

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD–MM–YYYY)* 27-03-2014	**2. REPORT TYPE** Master's thesis	**3. DATES COVERED** *(From — To)* Sep 2012–Mar 2014

4. TITLE AND SUBTITLE About Phase: Synthetic Aperture Radar and the Phase Retrieval Problem	**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Nelson, Aaron A, 2d Lt, USAF	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Air Force Institute of Technology Graduate School of Engineering and Management (EN) 2950 Hobson Way Wright-Patterson AFB OH 45433-7765	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT-ENC-14-M-03
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Intentionally Left Blank	**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Synthetic aperture radar (SAR) uses relative motion to produce fine resolution images from microwave frequencies and is a useful tool for regular monitoring and mapping applications. Unfortunately, if target distance is estimated poorly, then phase errors are incurred in the data, producing a blurry reconstruction of the image. In this thesis, we introduce a new multistatic methodology for determining these phase errors from interferometry-inspired combinations of signals. To motivate this, we first consider a more general problem called phase retrieval, in which a signal is reconstructed from linear measurements whose phases are either unreliable or unavailable. We make significant theoretical progress on the phase retrieval problem, to include characterizing injectivity in the complex case, devising the theory of almost injectivity, and performing a stability analysis. We then apply certain ideas from phase retrieval to resolve phase errors in SAR. Specifically, we use bistatic techniques to measure relative phases, and then we apply a graph-theoretic phase retrieval algorithm to recover the phase errors. We conclude by devising an image reconstruction procedure based on this algorithm, and we provide simulations that demonstrate stability to noise.

15. SUBJECT TERMS

Synthetic aperture radar, phase retrieval, angular synchronization, phase errors, circulant graphs, informationally complete, quantum mechanics, unit norm tight frames, computational complexity, Cramer-Rao lower bound

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dustin G. Mixon, Capt, USAF, PhD
U	U	U	UU	162	**19b. TELEPHONE NUMBER** *(include area code)* (937) 255-3636 x4516 dustin.mixon@afit.edu